UNIVERSITY OF MAINE
LIBRARY

Class no. 311
Book no. Su83
Accession no. 79,617

LIBRARY
UNIV OF MAINE

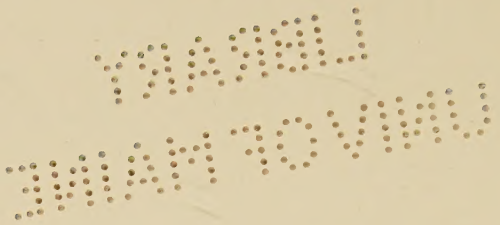
LIBRARY
UNIVERSITY OF MAINE

LIBRARY
UNIV OF MAINE

LIBRARY
UNIV. OF MAINE

LIBRARY
UNIV OF MAINE

ELEMENTARY STATISTICAL METHODS



McGraw-Hill Book Co. Inc.

PUBLISHERS OF BOOKS FOR

Electrical World ▽ Engineering News-Record
Power ▽ Engineering and Mining Journal-Press
Chemical and Metallurgical Engineering
Electric Railway Journal ▽ Coal Age
American Machinist ▽ Ingenieria Internacional
Electrical Merchandising ▽ BusTransportation
Journal of Electricity and Western Industry
Industrial Engineer

ELEMENTARY STATISTICAL METHODS

BY

WILLIAM G. SUTCLIFFE, A.M.

Assistant Professor of Economics
in Simmons College

First Edition

McGRAW-HILL BOOK COMPANY, Inc.
NEW YORK: 370 SEVENTH AVENUE
LONDON: 6 & 8 BOUVERIE ST., E. C. 4
1925

Copyright, 1925, by the
McGraw-Hill Book Company, Inc.

PRINTED IN THE UNITED STATES OF AMERICA

THE MAPLE PRESS COMPANY, YORK, PA.

Dedicated
To My Son

19,617

PREFACE

In the last few years an increasing demand for quantitative analysis in the business sciences has been developed. There is a constant search for statistical control, and this necessarily involves the development and use of statistical methods. This text is intended to furnish, to the college student and to all others interested in the collection, analysis, and presentation of economic statistical data, the elementary methods involved in such control. The author has reduced to a minimum the mathematical derivation of formulae, and has endeavored to present the subject without resort to the use of higher mathematics. By so doing, and by the use of synthetic series, it is to be hoped that the reader will acquire a better understanding and appreciation of the types of analyses rather than types of formulae.

In the arrangement of the text it will be noticed that the analyses of frequency and time series have been segregated into two distinct parts. This should make it possible for the reader to see the coordination or significance of all analyses pertaining to any one type of series.

It is intended that, if desired, the problems which follow many of the chapters should serve as laboratory exercises.

The author makes no claim to originality in the types of analyses given. He has endeavored to embody and explain such standard practices in such a way as to facilitate ease in computation and understanding. Numerous textbooks have been used for suggestions, and where quotations, data, and ideas have been used, an effort has been made to acknowledge the source.

The references given at the end of the chapters are by no means exhaustive—several books have been published since this book was written—but should serve as collateral readings on the topics discussed.

The author wishes to acknowledge his personal sense of indebtedness to Prof. T. H. Boggs of the University of British Columbia, who first stimulated his interest in economics, and to Prof. E. E. Day of the University of Michigan, who introduced him to the subject matter of statistics. To Prof. Sara Stites of Simmons

College, Alfred Boden the Town Clerk of Digby, N. S., Mr. A. Turnbull of the New Brunswick Telephone Company, St. John, N. B., Miss N. A. Sutcliffe, and finally to his wife, the writer is under obligation for many helpful suggestions in the compilation of this text.

William G. Sutcliffe.

Simmons College,
Boston, Mass.,
September, 1925.

CONTENTS

PAGE

PREFACE vii

PART I

GENERAL STATISTICAL METHODS 3–87

CHAPTER I

STATISTICS AS A SCIENCE. 3–6
1. Brief early history 3–5
2. Statistics as a science. 5
3. Definition of statistics 5
4. Definition of statistical methods 6

CHAPTER II

THE NEED FOR STATISTICS. 7–15
5. Introduction. 7
6. Statistics not mere figures. 7–8
7. The service of statistics to the government 8–12
8. The service of statistics to business. 12–15
9. Summary. 15

CHAPTER III

PROBLEMS PRIOR TO COLLECTION OF DATA. 16–22
10. Need for well-defined purpose 16
11. Definition of units of measurement. 16–18
12. Sufficiency of and kinds of data 18–19
13. Standards to be applied to secondary data. 19–21
a. Homogeneity of conditions.
b. Absence of bias.
c. Accuracy in tabulation.
14. Summary. 21
Problems 22

CHAPTER IV

THE COLLECTION OF PRIMARY DATA. 23–31
15. Introduction. 23
16. Methods of collecting primary data. 23–26
a. Transcription of records.
b. Enumeration.
c. Estimation.

PAGE

17. The schedule 26–29
a. The form.
b. Type of questions.
18. Instructions in the gathering of material 29–30
a. To enumerators.
b. To informants.
19. Summary 30
Problems 31

CHAPTER V

CLASSIFICATION AND TABULATION 32–49
I. Introduction 32
II. Classification 32–36
20. Definition 32
21. Mechanical aids in classifying data 33–34
22. Care in editing data 35–36
a. Verification of abnormal.
b. Accuracy of statement.
c. Consistency.
23. Summary 36
III. Tabulation 36–48
24. Kinds of tables 36–39
a. General-purpose.
b. Special-purpose.
25. Types of series 39–42
a. Temporal.
b. Spatial.
c. Component.
d. Frequency.
26. The methods of tabulation 42–48
a. Title.
b. Stub and captions.
c. What should be stub and caption.
d. Arrangement of stub and captions.
e. Rulings.
f. Aids in reading.
IV. Summary 48
Problems 49

CHAPTER VI

GRAPHIC METHODS 50–87
I. Introduction 50
II. Diagrams and maps 50–52
27. Defects of area diagrams 46–47
28. Secrist's standards for graphic presentation 52
29. Résumé of methods for illustrating different series . . . 53

PAGE

30. Pie diagrams 53–56
31. One-hundred per cent bar diagrams 56
    a. Simple.
    b. Composite.
32. Bar diagrams 57–61
    a. Need for horizontal zero.
    b. Composite bar.
    c. Vertical bar.
33. Maps 61–64
    a. Types in general.
    b. Static maps.
        (1) Cross-hatched.
        (2) Dot maps.
        (3) Limitation of static map.
    c. Dynamic maps.

III. Curves 64–87
34. Introduction 64
35. Determination of scales 64–7
    a. Field of chart.
    b. Vertical zero.
36. Methods of indicating vertical zero 67
37. Placement of figures about axes 69
38. Rulings for axes 70–71
39. Interpretation of curves ruled on rectilinear paper . . . 71–72
40. Superposition of curves 72–75
    a. Percentage charts.
    b. Scale adjustments.
41. Zone curve 76
42. Band diagram 76
43. Logarithmic charts 76–84
    a. Description.
    b. Why used.
    c. Derivation of logarithmic rulings.
    d. Contrasted with rectilinear paper.
    e. Interpretation of slope on logarithmic paper.
44. Summary 84
    Problems 85–87

PART II

STATISTICAL METHODS FOR FREQUENCY SERIES 91–225

CHAPTER VII

CLASSIFICATION AND TABULATION 91–111
45. Introduction 91
46. Definition of frequency table 91–92
47. An array 92–95

PAGE

48. Continuous series 95–97
49. Discrete series 97–98
50. The class interval 98–104
51. The class limits 104–108
Problems 108–111

CHAPTER VIII

GRAPHIC PRESENTATION OF FREQUENCY SERIES 112–132
52. Introduction 112
53. Frequency polygon or curve 112–114
54. Discrete series 114–115
55. Staircase curve 115
56. Smoothing curve 116–117
57. Series with unequal intervals 117–121
a. Rectilinear projections.
b. Logarithmic.
58. Comparison of curves 121–122
59. Types of curves 122–124
60. The Ogive 124–126
61. Interpolation by means of ogive 126–129
62. Lorenz curve 129–131
Problems 131–132

CHAPTER IX

METHODS OF ANALYZING A FREQUENCY SERIES 133–137
63. Introduction 133
64. Types of analysis 133–135
a. Measures of position.
b. Measures of dispersion.
c. Measures of skewness.
65. Types of averages 135–137
66. The desired attributes of an average 137

CHAPTER X

THE MEAN 138–148
I. The arithmetic mean 138–145
67. Introduction 138
68. Simple method of computation 138–140
69. The interval method 140–141
70. Points to consider in the use of interval method 141–143
71. The weighted arithmetic mean 143
72. Advantages and limitations of the arithmetic mean . . . 144–145
II. The geometric mean 145–147
III. The harmonic mean 147–148
Problems 148

CHAPTER XI

PAGE

THE MEDIAN, QUARTILES, AND PERCENTILES 149–155
I. The Median 149–153
73. Introduction 149
74. Computation of the median 150–152
75. Advantages and limitations of the median 152–153
II. Quartiles, deciles, and percentiles 153–155
76. Introduction 153
77. Method of computation 153–155
Problems 155

CHAPTER XII

THE MODE 156–159
78. Introduction 156
79. Moving average 156–157
80. Interpolation 157–158
81. Pearsonian formula 158
82. Advantages and limitations of the mode 158–159
Problems 159

CHAPTER XIII

SUMMARY ON AVERAGES 160–162

CHAPTER XIV

DISPERSION 163 171
83. Introduction 163
84. The range 165
85. Measures of dispersion 165
a. The mean deviation.
b. The standard deviation.
c. The quartile range.
86. The mean deviation 165–171
a. The standard to use in computation.
b. Actual deviation method.
c. The interval method.

CHAPTER XV

DISPERSION (*Continued*) 172–184
I. The standard deviation 172–175
87. Definition 172
88. Computation. Actual deviation method 172–174
89. The interval method 174–175
II. Properties and comparison of mean and standard deviation. 175–178
III. Coefficient of dispersion 178–180
IV. Quartile deviation 180–182
90. Definition 180
91. Measures of quartile deviation.

PAGE

92. The quartile coefficient 181
93. Advantages and limitations of the quartile deviation 181–182
V. Statistical résumé 182–183
Problems 184

CHAPTER XVI

SKEWNESS 185–191
94. Introduction 185
95. Definition 185
96. The need for such a measure 185–187
97. The measure of skewness 187–190
a. Foundation for measure.
b. Mean − mode method.
c. Quartile measure.
98. Coefficients of skewness 190–191
a. Mean − Mode measure.
b. Quartile measure.
99. Summary 191
Problems 191

CHAPTER XVII

CORRELATION 192–207
100. Résumé of preceding chapters 192
101. Correlation and regression 193
102. History of coefficient of correlation 193
103. Definition 193–195
104. The correlation table 195–198
105. How to measure correlation by means of standard deviation measures 198–202
106. Computation by actual deviations from the mean 202–206

CHAPTER XVIII

CORRELATION (*Continued*) 207–225
107. Interval method of computation 207–211
108. Formulae 211
109. Derivation of correlation coefficients by graphic methods. 211–215
110. Limitations of the coefficient of correlation 215–218
111. The probable error 218
112. The coefficient and lines of regression 218–223
113. Graphic presentation of lines of regression 223
114. Summary 224
Problems 225

PART III

STATISTICAL METHODS FOR TIME SERIES 229–330

CHAPTER XIX

TIME SERIES, THEIR PROBLEMS 229–230

CHAPTER XX

PAGE
INDEX NUMBERS 231–254
115. Definition 231
116. Brief early history 231–233
117. Summary of problems involved in the construction of index numbers 233–234
118. Types of data to be included in index 234–240
119. Selection and number of samples to use 240–248
a. Types of samples.
b. Number of samples.
120. Weighting 248–253
a. Introduction.
b. The system of weights to use.
c. Summary.
121. Summary of chapter 253–254

CHAPTER XXI

INDEX NUMBERS (*Continued*) 255–276
I. Chain relatives and fixed-base indices 255–262
122. Introduction 255
123. Chain relatives 256
124. The fixed base 257–263
a. The length of the base period.
b. The time of the period chosen.
c. Difficulties of changing base or commodities in a fixed-base index.
d. Summary.
II. The aggregative 263–268
125. Introduction 263
126. Aggregatives shifted to any base 263–265
127. Weighting and aggregative 265–267
128. Summary on aggregates 268
III. Averages to use 268–273
129. Introduction 268
130. Aggregatives may be devoid of averages 269
131. The geometric mean 269
132. Median and mode 270
133. The arithmetic mean 270
134. Averages to use 271–272
135. The ideal formula 272–273
IV. Summary 274
Problems 275–276

CHAPTER XXII

TIME SERIES 277–279
136. Introduction 277

PAGE

137. Types of fluctuations 277–278
a. Seasonal.
b. Secular trend.
c. Cyclical.
138. Summary 279

CHAPTER XXIII

THE PROBLEM TO WHICH ANALYSIS IS APPLIED. 280–282

CHAPTER XXIV

SEASONAL VARIATION 283–293
139. Introduction. 283
140. Causes for monthly variations. 284
141. Methods in general. 284–286
142. Link relative method. 286–292
a. Introduction.
b. Monthly relatives.
c. The representative relative.
d. Relative to January base.
e. Correction for secular trend.
f. Indices from mean of corrected indices.
g. Logarithmic derivation.
h. Graphic presentation.
143. Summary. 292–293
Problems 293

CHAPTER XXV

SECULAR TREND 294–307
144. Introduction. 294
145. Secular trend as a forecaster. 295
146. The measurement of secular trend 296
147. Graphic methods. 297
148. Secular trend by moving average. 297–299
149. Method of least squares. 299
150. Limitations of method of least squares 306
151. Summary. 306
Problems 307

CHAPTER XXVI

CYCLIC FLUCTUATIONS. 308–319
152. Introduction. 308
153. Normal production. 309
154. Moving average method. 309–310
a. Yearly totals.
b. Monthly.
155. Method of least squares. 310–311
156. Application of methods to the problem 311–317
157. Comparability of cyclic changes 317

PAGE
158. The group index 318
159. Summary 318
Problems 319

CHAPTER XXVII

BUSINESS BAROMETERS 320–330
160. Introduction 320
161. The effects of the cycle 321
162. The need and uses of a business barometer 321
163. The construction of a barometer 321–325
a. Its basis.
b. The grouping of series having similarity in time changes.
(1) Comparability of series by means of standard deviation.
(2) Grouping series which synchronize in time element.
(3) Correlation of time series.
(4) Lag.
c. Divergence of time changes between groups.
(1) Divergence between groups.
(2) The group index.
(3) Forecasting.
164. Summary 328
Problems 329–330

INDEX 331

PART I

GENERAL STATISTICAL METHODS

ELEMENTARY STATISTICAL METHODS

CHAPTER I

STATISTICS AS A SCIENCE

1. Brief Early History.—Scientific statistical analysis of data is of comparatively recent origin and it is an attempt to measure, quantitatively, phenomena encountered in the study of the natural and social sciences. It is true that statistics, as a science, are the direct product of the Industrial Age; nevertheless, statistics as mere compilation of figures antedate by many centuries, this era.

The Latin origin, *status*, of the term "statistics" suggests that, in the preindustrial period, statistics pertained chiefly to the state. Census returns were known to exist in the period of the Roman Empire. To students of English history the statistical contents of the Doomsday Book of 1086 are well known. Within its two volumes were recorded the acreage owned and the valuation of all land within the realm, together with much supplementary information that could be used for administrative purposes. The enumeration or count of various phenomena has, no doubt, continued since the beginning of tribal society—hence statistical compilations are not of recent origin. These, however, were used largely for administrative purposes, and not for the investigation and formulation of new policies. Herein lies the difference between the statistical methods used prior to, and after, 1800.

During the sixteenth and seventeenth centuries considerable attention was given to the problems connected with public finance, and in the various treatises written during this period there is an attempt, though meager, to use statistics as a basis for the investigation and formulation of new policies. In England, by the close of the seventeenth century, considerable interest was taken in "political arithmetik," as statistics were

then called. This interest is evidenced by such works as "Natural and Political Annotations Made upon the Bills of Mortality," which was the result of an investigation, conducted by Capt. John Graunt of London in 1666, into the mortality and birth rates. In this study was indicated, for the first time, the now well-known phenomenon of the greater frequency of male births over those of female. Seventeen years later, Sir William Petty, who is considered by some as the founder of the mathematical school of economists, published his "Five Essays in Political Arithmetick." In these essays and other writings he emphasized the need for statistics in all the various activities of the state. Edmund Halley, the discoverer of the comet of that name, in 1693, aided the study of demography by his "Estimate of the Degrees of Mortality of Mankind." In this publication he built up a life-expectancy table, the idea of which forms the fundamental basis of the present life insurance system. In spite of all these efforts the terms "statistics" was not used until 1749, when G. Achenwall, a German statistician, employed it in his book, "Staatsverfassung der heutigen vernehmsten europaischen Reiche."

Notwithstanding the advances made in the use and the understanding of statistics during the seventeenth century, they were not used in the modern sense until after 1761, when a Prussian clergyman, J. B. Sussmilch, tried by means of arithmetic devices to prove the validity of the "natural order," and thereby develop a general theory of society. In his work he combined the descriptive function with the analytical process of statistics, using statistics both as a basis for action as well as a record of the past. The advent of this work marks the beginning of the period in which statistics (as figures) were developed into a systematic study or science.

Knies, Adolf Wagner, Quetelet, Hilebrand, Block, and Engel contributed to the early development, while A. Meitzen, E. Thorndike, Sir Francis Galton, Karl Pearson, Francis Edgeworth, Jacques Bertillon, C. Davenport, R. H. Hooker, G. U. Yule, and A. L. Bowley are the more recent contributors to the science. Since it is not the purpose of this chapter to give a history of the development of statistics, but merely to show that it is an old subject, the facts of which have been systematized; the reader is referred to other sources for information on the history of statistics. Such books as "The History of Statistics,"

edited by John Koren, and "History, Theory and Technique of Statistics," by A. Meitzen, should be consulted.

2. Statistics as a Science.—Statistics as a method has been developed in all fields of activity and is making its contribution to practically all the sciences; while, on the other hand, some sciences have contributed much in the development of statistical method itself. Biology has contributed much both to terminology and to method. The field for statistics has so expanded that one author has said: "The science of statistics is the chief instrumentality through which the progress of civilization is now measured, and by which its development hereafter will be largely controlled."[1] This striking statement will introduce the student to two important facts: first, the importance of statistics; second, that the study of statistics is a science.

3. Definition of Statistics.—A science is systematized knowledge and, therefore, statistics require system in collection and analysis. Figures devoid of purpose, unrelated to any particular subject, and gathered indiscriminately *cannot, therefore, be called statistics*, according to the meaning of the term as it is to be used in this text. *Statistics comprise the collection, tabulation, presentation, and analysis of a sufficient aggregate of facts, collected in a methodical manner, without bias and related to a predetermined purpose.* This definition is somewhat lengthy, but the qualifying clauses are necessary in order that statistics may be differentiated from mere figures. It also emphasizes the care with which statistics must be compiled and used. Science requires the search for truth; and truth cannot be found when prejudice or bias intervenes or when the data examined are a negligible quantity and not representative of the conditions they are supposed to represent. Conditions, likewise, must be related in some definite way; while it is impossible in the social sciences to isolate causes, a certain degree of homogeneity may be obtained in the conditions. All these facts should make apparent the need for the qualifying phrases in the above definition.

4. Definition of Statistical Methods.—*Statistical methods are all those devices used in the analysis and collection of statistics.* They embody all the methods of presentation and analysis that bring out the relationships existing within or between groups of statistics. Obviously, then, the care exercised in choosing and using these methods must be as great as in collecting the data.

[1] "History of Statistics," p. 14.

Part I of this book will consider the need for statistics and the elementary methods used in the collection, tabulation, and graphic presentation of statistics.

References

Bailey, W. B.: "Modern Social Conditions," pp. 1–14.
Bowley, A. L.: "Elements of Statistics," chap. I.
Meitzen, A.: "History, Theory and Technique of Statistics," pp. 3–54.
Meitzen, John: "Geschicte der Statistik."
King, W. I.: "Elements of Statistical Method," chap. I.
Koren, John: "The History of Statistics."
Pearson, Karl: "The Grammar of Science," Intro.
Secrist, H.: "An Introduction to Statistical Methods."
——: "Readings and Problems in Statistical Methods," pp. 1–26.
Yule, G. U.: "An Introduction to the Theory of Statistics," Intro., 5th ed.

CHAPTER II

THE NEED FOR STATISTICS

5. Introduction.—Statistics are regarded by some with opprobrium and the saying "statistics won't lie but liars will figure" has become almost proverbial. Others agree with Mark Twain that lies may be divided into three classes: "lies, damned lies, and statistics." These two statements are not consistent, but do imply that statistics may be used to prove anything. This is quite true, but the admission only emphasizes the need for statistical methods, and a knowledge of such methods, in order that the two statements may be rendered utterly false.

6. Statistics Not Mere Figures.—Statistics as mere figures have little utility; but when collected in sufficient quantity, without bias, for a predetermined purpose, and relating to a definite subject, they may be regarded as a lighthouse which indicates to the ship of state the right course to pursue. The efficiency of the lighthouse depends upon its location, the candle power of the light, and the character and state of its lenses. If the lenses are clouded by personal bias, the candle power weakened through paucity of data, and the structure erected too far from the conditions it is trying to indicate, the ship piles up with injurious results—statistics have lied or liars have figured. When a lighthouse proves worthless through faulty construction, society does not condemn lighthouses in general, for it knows that there are effective ones whose rays penetrate the thickest fog and indicate to the mariner the right course. Then why condemn statistics? Statistics should, and do, perform a similar function for society. When compiled with accuracy to all the conditions involved they *indicate* the channels to be followed; it remains, however, for society to see the light and take its bearings.

In spite of the antipathy of some towards statistics, statistics are used by practically everyone; unfortunately, however, the methods are not standardized. As John Koren states:

> We cry aloud for facts; there is a voracious and indiscriminating appetite for figures, or rather the nourishment they afford to argument

and propaganda: statesmen, teachers, preachers, publicists, and men in the street exemplify it. It is a dyspeptic appetite, if you please, because of the ill-sorted wares upon which it feeds.[1]

The retailer in selling shoes carries a smaller stock of extreme sizes; the small grocer in serving a poor community stocks few luxuries; each by studying the wants of his customers stocks accordingly, so as to avoid a large stock on hand which may burden the shelves for years. The housewife, in these days of balanced rations, is not immune from statistics, but draws up a caloric chart for the household consumption. The political uses to which figures may be put are only too well known. *Figures* have encroached upon every field of activity, but rarely is a thought given to statistical method. Effective work requires effective tools, and the workman should know the limitations and uses of his tools. Statistics are tools which, unfortunately, receive little attention from the user.

If figures are important to society, they should be used in a methodical way and utilized in the formulation of new policies. The thermometer in the hands of the physician serves as an indicator of general health conditions—its readings form the basis for action. The thermometer is of a special type, its calibrations have been accurately determined, in a methodical way. Statistics should be the social thermometer, with its calibrations determined as methodically as the physician's. Too many persons, unfortunately, have been using the kitchen thermometer to measure a social temperature. What is needed more than anything else in statistics is an understanding of methods and the standardization of such methods. The more universal the understanding of the methods and standards involved the less is the possibility for liars to figure and "get away with it," and for statistics to lie. Statistical control requires statistical methods.

7. The Service of Statistics to the Government.—In emphasizing the need for method it would be well to indicate the need for, and the services of, statistics. Statistical needs seem to be ever widening; the systems of production, consumption, distribution, and exchange are no longer considered omniscient, but are capable of, and demand, revision. The revisions should be based upon the records of past achievements and failures as

[1] KOREN, JOHN, "Some Statistical Standards," *Am. Stat. Pub.*, p. 351, March, 1915.

they relate themselves to the present. Social advance implies social regulation, which is supposed to be conducive to the welfare of all; legislation should be based upon facts.

> Our proneness to reason and act, through legislation and otherwise, upon insufficient fact bases, to make a fetish of theories without subjecting their value to the simplest test, do not make for intelligent progress, no matter what political or social problem is involved.[1]

The truth of this statement must be apparent to all who read the mass of Federal legislation proposed (not always passed) as sop for the farmers.

Unrestricted immigration was regarded by most people —and still is by a few—as a boon to society. Statistical research, though meager, finally indicated that restriction was necessary if the American standard of living was to be maintained. Restriction was written into the Immigration Law, and now in 1925 the battle rages over the type of restriction—whether or not there is Nordic superiority, and the necessity for Japanese exclusion. A statistical thermometer is needed to measure the physical, mental, and moral qualities of the types of immigrants seeking admission to the United States. Very little has been done in this field; studies have been made of the ability of immigrants to speak English, or the nationality of the inmates of the various state and Federal institutions, but hardly sufficient evidence has been deduced to indicate the ratio of exclusion. Statistics should be of great service in the framing of legislation; unfortunately, legislation is usually written first, and the conditions which it aims to remedy are investigated afterward.

The National Bureau of Economic Research, working in conjunction with the Conference on Unemployment held in September, 1921, has issued a statistical study on unemployment, and the facts therein collected should form a basis for future action. The Federal Coal Commission has recently made a study of the bituminous and anthracite industries and reports that there are too many mines in operation in the soft-coal fields; the report emphasizes statistical control of production.

Amalgamations of competing concerns were frowned upon as being detrimental to society, with the result that blanket antitrust laws were passed. The Interstate Commerce Commission, however, after investigating railroad conditions, advocated

[1] *Ibid.*, p. 355.

in the Transportation Act of 1920 the need for the consolidation of some competing lines.

Taxation is a method through which the state obtains much of its revenue. Statistics, and not sentiment, should be the basis of rate determinations. High rates, it is said, drive capital into tax-exempt securities (if the tax be the income tax), or make it resort to methods of evasion, so that the revenue derived under a high rate tends to be smaller than if the rate were lower. Statistics are necessary to prove the validity of these statements and an enlightened tax system should make use of such methods.

> Taxation is peculiarly a field cursed by non-expert opinion and unenlightened legislative action. I know of no field of statistical inquiry yielding more positive results or where honest investigation is more needed.[1]

Morbidity statistics are of extreme importance. How is it possible to measure the effectiveness of special treatments of disease if there are no records of the cases and conditions of the patients? In this field, standardization of method of, and compulsion in, reporting cases of disease is one of the pressing needs of the moment. Vital statistics present similar problems; the registration area should include all territories within the United States.

Occupational diseases, disability statistics, are very important from the social point of view; statistics can and do serve in this connection. Royal Meeker, writing in the *Monthly Labor Review* for March, 1920, states:

> The heavy toll of deaths due to blast-furnace accidents led to improvements in the construction by which "blow outs" with their accompanying fatalities have been almost entirely eliminated. The substitution of mechanical charging of blast furnaces for the old, hand-charging methods has practically eliminated fatal "gassing" among the men who were obliged to work on the top of the old type blast furnaces. These improvements in the construction of blast furnaces were brought about as the result of analyses of accident statistics and the calling of the attention of those responsible for the direction of industry to the causes of the more severe accidents . . .
>
> The men who are engaged in analyzing accident statistics by tabulating them by causes and severity of injuries are like the general staff

[1] James, A. E., "Measures of Relative Tax Burdens," *Am. Stat. Pub.*, March, 1916.

of the army. They map out the objectives of the campaign and furnish the information necessary to attain these objectives most speedily and economically.

The above quotation indicates how statistics have aided and served as a lighthouse by indicating the channel to follow, but even in this field perfection has not been reached, as the quotation continues to state:

Let the watchword be the prevention of preventable accidents. To attain this goal, accurate, complete, and comparable accident statistics are indispensable. Better statistics make for better safety work.

As an illustration of a law passed through statistical investigation, the Act of 1913 passed by the Massachusetts Legislature which regulated municipal finance may be cited. The Act aimed at the abolition of the well-known evils in municipal finance:

First, incurrence of funded or fixed debt for current expenses; second, temporary borrowings to an unlimited amount in anticipation of tax collections; third, diversion of the principal of trust funds to current expenses for unauthorized objects, and incurrence of other liabilities without proper provision for payment; fourth, neglectful and costly management of sinking funds.

This legislation is a notable example of scientific lawmaking by application of the statistical method. The director of the Bureau, Mr. Charles F. Gettemy, who deserves great credit for the success of this experimental undertaking in state investigation and supervision of municipal finances, well says: "The true method of undertaking to accomplish a reform of long-standing evils should be, in my judgment, by a preliminary, scientific diagnosis of conditions, and I doubt whether any state can show an instance of legislation more thoroughly predigested than that which has just been placed upon the statute books of Massachusetts for the purpose of restricting and regulating the incurrence of municipal indebtedness in accordance with sound financial principles.[1]

Contrast the above statement with the type of law which has been passed to help the farmer. Laws based upon sentiment or politics usually become dead letters or require other legislation to remedy the defects of the former. An outstanding example of the formulation of laws or treaties can be found in the reparation situation. After six years of action, reaction, sentiment, and

[1] Baldwin, F. S., "Statistics in the Service of the Municipality," *Amer. Stat. Pub.*, p. 106, June, 1914.

politics, people finally realized that a statistical study made on a scientific basis would be the only way to arrive at the indemnity payments for the Great War—hence the Dawes Report. Statistics, however, like the lighthouse, merely *indicate* the channel to follow, society may take its bearings or it may be so blinded by sentiment and passion as to ignore the light and eventually pile society up on the rocks of financial ruin.

8. The Service of Statistics to Business.—Many more illustrations may be given of governmental activities which require statistical indicators; the above are only a few from a great wealth of subjects, but they are sufficient to make apparent the importance of statistics to the state. If statistics are important to the group as a whole they may also be utilized by the component parts of the whole. Statistical control of industry is of extreme importance. Organized labor, once so hostile to business statistics because of the bias with which they were usually collected, now maintains its own statistical bureau. Business is becoming a science, and with the multiplicity of the factors entering into market conditions the competing industries, in order to survive, must have statistical control.

Statistics should serve industry by indicating:

1. The type and quantity of goods to stock.
2. When to buy and push sales.
3. Advertising mediums to use and the results therefrom.
4. The relation of the particular industry to the business cycle.
5. Production and costs.
6. Wages and labor cost per unit. The saving on labor-saving devices.
7. Cost of labor turnover.
8. Operating ratios.
9. Seasonal character of the business.
10. Credit conditions.
11. Efficacy of various methods of industrial relations, with employees and other concerns.
12. Tariffs and trade with foreign countries.

All these topics and others are of vital concern to the business man, and lack of knowledge may prove costly. Today the larger concerns have established special departments for statistical research because they realize many problems may be

treated in a scientific way. There is no "rule of thumb" which may be applied to business in general, but every concern has its own particular problems and they should be dealt with in a scientific manner.

The phenomena of the business cycle are familiar to students and the dislocation wrought in industry through the period of depression in 1920 and 1921 was felt by all. The cycle seems almost inevitable, but the oscillations within the cycle appear to be capable of modification. Statistical services, such as Babson's, Brookmire's, Moody's, and the *Review of Economic Statistics*, devote their energies to forecasting future conditions, so that the business man may have time to set his house in order and prevent the crests and troughs of the cycle being too pronounced. Individual business concerns find that they have a particular relationship to the business cycle, and, by statistical studies used in conjunction with the forecasting services, they are able to regularize their industries. The president of one large manufacturing concern states:

> So we discovered that the second suggestion of the economists also found application in our practice. We not only had a business cycle of our own, but our own cycle had a definite measurable relation to that of general business. And we were, therefore, in a position to use in a very tangible way the forecasts which constantly passed through the Planning and Statistics Section.[1]

Summarizing the results, the article continues:

> Of all these benefits, probably the greatest has been our ability to operate on an even basis throughout the year. Under the old conditions we would have produced 16 per cent of our total requirements in the first quarter; 22 per cent in the second quarter; 28 per cent in the third quarter; and 34 per cent in the fourth quarter. In 1922 it was possible to equalize approximately the quarterly production figures. In these days when seasonal employment is one of the country's greatest problems, when labor turnover is great, and when actual shortage of labor has occurred, continuous employment at fair wages is the stimulus to the good workman that no temporary attraction can equal.

The student by glancing at the contents of business magazines will be convinced that statistics are being used. Articles such as "Six Plans to Increase Sales," and "Seven Plans That Cut Our

[1] COONLEY, H., "The Control of an Industry in the Business Cycle," *Harvard Bus. Rev.*, July, 1923.

Buying Expenses" indicate that statistics must be used in measuring the efficacy of these methods.

A brief consideration of the costs of advertising should convince the student that money is not spent without some estimate being made of the possible returns. Advertising campaigns are conducted on scientific lines; policies must conform to local conditions with regard to class stratification and climatic conditions. A thorough knowledge of the field is of extreme importance and the preliminaries to any campaign may consume much time and analysis. H. K. McCann states:

> This search for knowledge leads into all sorts of paths and may consume months of time. It may include the examining of government and trade reports and statistics; the delving into history for general background material; even scientific experiment in medical or food lines for example, accurate first-hand contact with enough of the people who will sell and buy your product to give a fair picture of trade conditions and consumer buying habits. Advertising is ultimately associated with actual sales work and conditions to be met in the market. With all the facts before us collected, tabulated and analyzed, we are equipped to prepare the advertising campaign.[1]

Because of the intricacies of business and the interrelations of industries, statistical planning enters into almost every phase of business activity. The planning consists not only in the compilation of data and reports to the central executive, but also in furnishing information to the executive with regard to policies, and initiating investigations. Statistics indicate the path to follow and have a very important function to perform, the performance of which depends upon the utilization of the proper methods and a realization of the limitations of such methods.

One author has summarized the service of statistics to economics as follows:

1. Afford an accurate form of summary statement of economic facts and changes.
2. To develop relationships of facts within particular groups of data, and lay bases for empirical laws.
3. Enable us to determine the direction or continuity of movement of the group with which they deal.
4. Serve as a basis for legislation and administration.

[1] McCann, H. K., "Planning and Preparation of an Advertising Campaign," *Harvard Bus. Rev.*, April, 1923.

5. The possibility of tracing the effects of economic legislation.
6. For the purpose of developing a general theory.[1]

9. Summary.—By this time the student should be fully convinced of the need for, and the uses of, statistics. With need must come the development of statistical method in order that the method of attack may be rational. Statistics, however, should not be considered as a panacea for all ills, for they have limitations which should be recognized. They are tools, and as such are only capable of certain applications, but with good craftsmanship—which implies judicious use—a structure may be reared which will act as a beacon not only to society in general but also to the individual. The development of the tools and their application have made the study of statistics a science. The following chapters will discuss some of the tools used in the collection and analysis of statistics.

References

Bowley, A. L.: "An Elementary Manual of Statistics," chap. I.
Coonley, H.: "The Control of an Industry in the Business Cycle," *Harvard Bus. Rev.*, July, 1923.
Copeland, M. T.: "The Business Statistics," pp. 98–115.
Falkner, R. P.: "The Scope of Business Statistics," *Quar. Pub. Am. Stat. Assoc.*, June, 1918.
King, W. I.: "Elements of Statistical Method," chap. III.
Koren, John: "The History of Statistics," pp. 15–49.
McCann, H. K.: "Planning and Preparation of an Advertising Campaign," *Harvard Bus. Rev.*, April, 1923.
Newsholme, A.: "Vital Statistics."
Parmalee, J. H.: "Public Service Statistics in the United States," *Quart. Pub. Amer. Stat. Assoc.*, June, 1915.
Secrist, H.: "Statistics in Business," pp. 98–115.
——: "An Introduction to Statistical Methods."
——: "Readings and Problems in Statistical Methods," pp. 22–46.
Quart. Pub. Amer. Stat. Assoc., vol. 14.

[1] Kinley, D., "The Service of Statistics to Economics," *Amer. Stat. Pub.*, pp. 113–120, March, 1914.

CHAPTER III

PROBLEMS PRIOR TO COLLECTION OF DATA

10. Need for a Well-defined Purpose.—No one would think of erecting a building without a carefully drawn plan, a plan which in itself would be determined, to a great extent, by the purpose of the building. That is to say, a definite idea of the purpose of the structure is absolutely necessary before measurements or detail are actually thought of—purpose dominates the plan. By means of statistics, structures giving different analyses are evolved which conform to a definite purpose. The first rule to be observed in starting a statistical problem is: *Have the problem definitely in mind, and know whether or not it will lend itself to mathematical treatment.*

As an illustration for the need of a well-determined purpose, suppose the investigator has only a vague idea that he wishes to find out the cost of production of a given article. He collects cost figures from all departments of the factory and then suddenly awakes to the fact that what is really desired is direct labor costs and not total costs. Because of the lack of a predetermined purpose the data collected are of no use—as they are expressed in the wrong unit of measurement, total costs instead of labor costs. The analysis of the figures already collected would lead to erroneous conclusions; the labor, therefore, has been misapplied. This may appear to be an absurd illustration, but investigations are constantly being made which are of little use, because the investigator started out to investigate something and did not take the trouble to find out definitely what that something was.

Statistics are tools, and as such they have to be chosen according to the type of structure to be reared. The good craftsman works to specifications, and the plan determines the order of the various processes, and these in turn indicate the time and place for the use of the different tools. In the last analysis *purpose governs all things.*

11. Definition of Units of Measurement.—It is evident from the previous illustration that purpose determines, to a great

extent, the unit of measurement. When the architect has decided upon the type of structure to be built he must draw up a plan, which includes all the specifications. This plan is reduced to scale, a fraction of an inch representing a foot. But if the building happens to be a large business block the scale —or fraction—will tend to be smaller than if the structure were a summer cottage. He has to determine the unit of measurement, and must rigidly define it in his blue prints, so that all plans for the same building—front elevation, floor plan, etc.—may be seen in the right perspective. Likewise, in statistics, all structures must be reared according to a definite plan drawn to a rigidly defined scale, so that all the data entering into the structure may be of the same unit of measurement. By this means the different angles of vision (averages, dispersion, etc.) may be given in the right perspective.

A building plan, however, may have one of several scales, and the unit used has a definite connotation, an inch always being the same distance, whereas a statistical plan may have units which vary considerably, and, furthermore, like terms may have totally different connotations. As already indicated, costs might mean one of many things—labor costs, overhead, and selling costs, etc. In order that no edifice may be constructed with unlike units of measurement the term "cost" must be rigidly defined. If the study involves direct labor cost the term "direct" must be defined so that the investigator may know *definitely* what materials to include or exclude from the data. The United States government is most painstaking in defining the terms used in its revenue bills. For the definitions of the terms used in the taking of the census the student is referred to the pamphlet issued by the Bureau of Census entitled "Instructions to Enumerators."

Accident statistics are familiar to most students of industry, yet how many realize the possible interpretations of the term "accident." As a general term it must be defined. What constitutes an accident—any injury whatsoever, or is it based upon duration or disability? The Statistical Committee of Workmen's Compensation has defined the term as those injuries which disable an employee for a period longer than the day of injury. This is the definition of a general term; very often information is required of specific accidents. These may have a general classification as follows:

1. Fatal accidents.
2. Those involving permanent disability.
3. Dismemberments.
4. Partial disability; duration and degree.
5. Temporary disability; kind and duration.

In each of these five classes the general term "accident" has been so qualified that the unit of measurement has changed. What does "fatality" imply? Does it include slow deaths in the form of occupational diseases? What does permanent disability imply? What time must elapse to differentiate this type of accident from temporary disability? As to "partial disability"—how is the degree of disability determined? As to "temporary disability" —"what is the time element? All these questions should be answered before the investigator begins to think of collecting data. In statistics, therefore, it is especially important that the second rule be given due consideration, which states: *The unit of measurement should be rigidly defined and there should be included in the study only those data which possess the specific type of measurement.*

12. Sufficiency of and Kinds of Data.—With purpose and plan in mind the builder now looks over the locality to see whether or not his type of building is needed, and, if so, to see if it is possible to get the necessary materials for its construction. So with statistics, the structure may have been erected by others. But if the problem has been done before, the data may be so scattered and unrepresentative that they would not suffice to make a worthy edifice. Hence the third rule follows from the others: *See if the problem has been done before, and, if not, investigate the quantity and quality of data available for the project.* Never start a study without having surveyed the field to determine whether or not the energy required to make the study might be wasted. Studies of a similar character may have been made before, but conditions may have changed so that a new study might be justified after the nature and the scope of the old have been investigated.

Satisfied as to plan, measurements, and availability of materials, the workmen are now ready to start on the foundation. Material begins to concentrate around the project and the work goes on according to specifications. The foundation, while not appealing to the eye, is of great concern to the engineer, for it

must bear all the stress and strain. The whole superstructure depends ultimately upon the nature of the foundation; therefore, it must be of the best materials. Usually, it is made of new material of known stress but sometimes the foundation includes old materials, before such are used, however, they are subjected to rigid examination for quality and fitness, so that only the very best may be utilized. In the finished structure the ingredients of the foundation are lost sight of, but the owner of the building—having faith in his contractor—has little to fear because he knows the character of his foundation.

Data are the foundation of statistical studies and, unfortunately, studies are made in which little consideration is given to the quality and the fitness of the materials used. When the contractor is not known the figures should be given the closest scrutiny; hence the rule: *Tables should indicate the source from which they were compiled or derived.*

Data may be divided into two classes:

1. Primary or new.
2. Secondary or derived.

The process of the selection of primary data will be taken up later. It is sufficient here to state that primary data are analogous to the new materials used in the foundation of a building. Secondary data are analogous to the old materials, which were collected by others at previous periods of time. A good foundation is absolutely necessary for a statistical study; therefore, the materials embodied in it must conform to certain standards.

13. Standards to Be Applied to Secondary Data.—It is not sufficient to embody secondary data in a new study because it bears the same name and was derived from approximately the same district. Remember that in statistics the same name may have different connotations. As an illustration, the investigation under way might be in the field of vital statistics, the marriage rate in district X. Let it be supposed that work has been done in this field before, by some other person, and that the previous investigation covered all the wards, except two, of this same district X. Because the two studies are marriage rates, and cover a part of the field in both investigations, it does not follow that the investigator needs only to canvass the two extra wards and add these data to those contained in the previous study.

The unit of measurement of the one study may be totally different from that of the other, even though they are both termed "marriage rates." The first investigation may have been undertaken for the purpose of comparing the marriage rate in district X at different periods of time, the rate being expressed as a number per 1,000 of the *total population.* The recent study may propose to show the comparison of the rate in district X with those in other districts. If such is the object, by realizing that some districts may be on the frontier where there would be relatively few children and aged people, while other districts, long since established, would have a fair representation of all ages, the unit used must be that which will express the rate per 1,000 not of total population but rather the *population of unmarried and of marriageable age.* Such a unit would consider widowed and divorced people as unmarried; the age limit for marriage would have to be determined upon, it being usually fifteen years. The units of the two investigations, while having the same spelling, are dissimilar, chiefly because the purpose of the investigations differed. The material collected in the earlier study cannot be embodied in the recent investigation. *Secondary data must be carefully analyzed to see if the units of measurement, and the purposes of the studies, are identical.*

Another test must be applied to secondary data. While they may conform to measurements and purpose, the time elapsing between the old study and the new may be so great that the conditions have changed. The type of architecture has changed; the frontier life has changed and is now a thing of the past. *Conditions under which the data were collected must, therefore, be homogeneous with those of the present.*

The secondary data may be found twisted and warped by the personal bias of the investigator, thus giving the data a weak and false foundation. Such material must be rejected, as its inclusion in any new study would undermine the new structure. At all times such data must be given the closest scrutiny to determine whether or not *they are free from bias.*

Bias may not be present in the data but, owing to their paucity or poorly chosen samples, or because of inaccuracies in tabulation, they cannot be considered as representative of the facts they endeavor to portray. Such considerations become important criteria for the inclusion or exclusion of secondary data in any problem.

Summarizing, it may be said that, to include secondary data as a part of any investigation, they should:

1. Be compiled in the same units of measurements and for the same purpose.
2. The conditions portrayed should be identical with those of the new study.
3. There should be no personal bias within the data.
4. Accuracy of tabulation and representativeness of facts should be assured.

Too much emphasis cannot be laid upon the care with which secondary data should be used. Remember, at all times, data are the foundation of statistics and determine, to a great extent, the efficacy of the structure reared upon that foundation.

Knowledge of the existence of secondary data is important. Such data may be found in the official records of the government—Federal, state, and municipal—chambers of commerce, health centers, and trade magazines. In many studies the data are purely local and the sources are well known. For more general studies, information as to sources may often be secured by consulting "The Source Book of Economic Data," by L. Haney, published by Prentice-Hall.

14. Summary.—Summarizing this chapter, it is only necessary to remind the student that statistics require just as careful planning as the construction of a building. The analogies used make this evident. Beautiful architecture avails little if superimposed upon a weak foundation, for the building will finally sag and crumble. Refined analysis of statistics is futile if the original data were compiled in a haphazard way. A carefully constructed plan, using the best of materials and craftsmanship, is the only guarantee that statistics will not lie.

References

Bowley, A. L.: "An Elementary Manual of Statistics," chap. VIII.
——: "Elements of Statistics," chap. II.
King, W. I.: "Elements of Statistical Method," chaps. IV–V.
Meitzen, A.: "Statistics," pp. 168–170.
Secrist, H.: "An Introduction to Statistical Methods."
——: "Readings and Problems in Statistical Methods," chap. III.
——: "Statistical Units as Standards," *Quart. Pub. Amer. Stat. Assoc.*, March, 1918.
Zizek, Franz: "Statistical Averages," pp. 25–33.

Problems

1. What would be the initial step taken by you in an attempt to measure the following?

a. The saving involved through the use of better shipping containers.

b. The relative merits of motor truck and rail transportation for the short haul.

c. Labor turnover in a given factory.

d. Changes in commodity prices.

e. Advertising costs.

f. Stock turnover.

g. The spread between the price paid to the farmer for a box of apples and the price paid by the consumer.

h. The growth in improved farm acreage.

i. The decrease in industrial accidents.

j. Waste in industry.

2. What would be the definition of the unit of measurement, for each of the above problems?

3. After the determination of the unit of measurement, would you, at once, begin to collect primary data? If so, why? If not, why not?

4. From what source would you obtain information for the problems above?

5. Suppose a large holding company of public utilities issued a financial report which caused heavy selling of the stock held by the small security holders and thus lowered security prices. The stock, however, immediately reacted and reached new high levels. If you were making a statistical study of the financial conditions of public utilities, would you include this company report in the data? If so, why? If not, why not?

6. Where would you expect to find data on the following general topics?

a. Coal production.

b. Immigration.

c. Bank clearings.

d. Unemployment.

e. Financial statistics.

f. Crop conditions.

g. Exports and imports.

CHAPTER IV

THE COLLECTION OF PRIMARY DATA

15. Introduction.—From the last chapter it will be recalled that a statistical problem, to be effective, requires a well-developed plan—one that has taken into consideration the units of measurement and the availability of the data for the special study. The problem which now demands attention pertains to the collection of primary data.

16. Methods of Collecting Primary Data.—Primary data are facts collected from original sources. They are the raw materials entering into the statistical structure, and the requirements for the collection of such data must be as rigid as those concerning the analysis of the sources and conditions of secondary data.

The methods of collecting primary data depend, largely, upon the following factors.

1. Nature of the problem. Size of the territory covered in investigation.
2. The funds available for the collection of data.
3. The powers of collecting data (mandatory or voluntary).

If an investigation involves some phase of industry for which adequate and homogeneous records are available, the material may be obtained by the simple process of the transcription of records. This may be done by agents, informants themselves, or by personal investigation, the method being determined according to the funds available and the size of the territory to be covered. As an illustration of such a problem, an investigation might be made of some phase of railroad administration. Since the Interstate Commerce Commission has compelled the railroads to adopt uniform accounting systems, the method used in securing data would be that of transcription, because all the data would be comparable. A secondary source for such a study would be the reports submitted by the railroads to the Commission, as prescribed by law.

The nature of the problem may be such that representative data would suffice—samples might be used instead of including

all the data. By such a method the time and the cost of collection would be much less. Such studies, nevertheless, require careful collection because the samples chosen must be representative.

If the territory is large, with available funds for investigation, together with mandatory powers for the collection of data, as possessed by the United States government in the taking of the census, the data are usually collected by paid agents called enumerators. This process is called collection by enumeration, since it involves the enumeration of all possible items. (For mandatory powers in taking the census, see "Instructions of Enumerators," already referred to.)

Some idea of the expense involved in taking the census of the United States may be realized when for the Thirteenth Census (1910) some $16,000,000 was spent, or a per capita cost of 17 cents; the estimates for the Fourteenth Census (1920) exceeded $20,000,000.[1] The magnitude of the operations is discussed by Director of Census, William Steuart, who states:

> The Fourteenth Census covers population, with special schedules for the blind and deaf; agriculture, with a special schedule for animals in barns and enclosures not on farms; manufactures (including forest products), with 85 supplemental schedules; mining, with 44 supplementals; irrigation and drainage. There are in all 138 schedules upon which data required for some branch of the census must be reported. There are hundreds of inquiries to be answered—not by every individual but some by many individuals.[2]

Speaking of the enumerators, he continues:

> . . . the applicants were required to prepare formal application, to make a practical test, and to attend schools of instruction; but there were more than 86,000 enumerators to be instructed, and the instructions were necessarily more or less superficial.

Realizing the cost and organization necessary for such nation-wide projects, it can readily be seen why private individuals can rarely use the enumeration process where a large territory is involved; hence in such fields the estimate or sample process is used.

[1] Hill, J. A., "The Essentials of a Good Census," *Amer. Stat. Pub.*, March, 1920.

[2] Steuart, William, "The Conduct of the Fourteenth Census," *Amer. Stat. Pub.*, p. 573, March, 1921.

If, on the other hand, the territory is large, the available funds for collection small, and the information dependent upon the volition of the informant, the data may be collected on the estimate basis. In such a method not all the data are collected but from those obtained is computed an estimate to cover the entire field which is supposed to be representative of the whole. The data may be obtained by having correspondents in certain sections of the field, by soliciting information through direct correspondence, and by the transcription of records. All these methods may be combined in one study, as was done in the two studies made by the National Bureau of Economic Research: "Income in the United States," and "Employment, Hours, Earnings in Prosperity and Depression." The student is referred to the first chapter of each book for a dicussion of the methods employed in soliciting the data.

Needless to say, much data secured by voluntary methods may contain certain inaccuracies, while other desired material may be omitted from the record. Such defects will always occur because the informant is skeptical of the use to which the information may be put. These discrepancies, however, may be reduced to a minimum by the use of a well-constructed schedule, which will inspire confidence in the informant.

Sometimes the inquiry is conducted throughout by the person interested, in which case the amount of funds is usually negligible as compared with the amount of time required. Such studies must be either purely local in character or conducted by the sample method. The time consumed in personal investigations may, in some protracted studies, be a disadvantage because the analysis, when derived, may not represent the actual conditions, as changes may have taken place after the investigation was begun. This rarely occurs, and when it does it indicates that the territory was too large for the personal method. Such a method has an advantage over the enumeration process for a small field, for the data should be more accurately recorded, assuming, of course, no personal bias. Personal investigations are largely local in character and most often are conducted by the school nurse, community health centers, business offices, etc. These inquiries usually have little effect upon national policies unless such results are considered as samples of larger groups, for example accident statistics, where a particular factory is considered representative of the group. The importance, how-

ever, of local investigations is hardly realized and should receive more attention from public-spirited citizens.

Summarizing, the nature of the problem dominates the field, and no problem, the nature of which requires more time and money than is available, should be undertaken. However, the type of problem cannot always dominate the power of collecting data. A problem to be relatively effective may have to resort to voluntary information, mandatory powers being prohibited. In spite of such a handicap, investigations have been made effective by a judiciously worded schedule.

Irrespective of the method used in the collection of data, a form is required, whether it be filled by informant, enumerator, or investigator, upon which the data must be recorded. These forms are called "schedules" and upon their structure depend, in a large measure, the amount and the reliability of the information, especially if mandatory powers are lacking. Considerable attention should, therefore, be given to the construction of the schedule.

17. The Schedule.—As an illustration of an extensive schedule the student is referred to the form used in taking the United States Census of Agriculture as given in the Census Report on Agriculture. For the other census schedules the student is referred to the various volumes of the census. As examples of forms lacking mandatory powers see "Employment, Hours, Earnings in Prosperity and Depression," by the National Bureau of Economic Research, pages 10–15.

Schedules cannot be constructed by any rule of thumb because, once more, purpose dominates the contents of the questionnaire. A few general rules, however, may be given which apply in a general way to the construction of forms:

1. The informant should be assured that the information solicited is confidential and its origin will not be divulged, but will be buried in the mass of data, unless it is a mandatory inquiry to see if any laws have been contravened, like many of the investigations conducted by the Federal Trade Commission. The agricultural schedule indicates the data are "confidential."

2. The informant should be told the purpose of the investigation or study, either in positive or in negative terms. The illustration places the object in the negative sense—"Not connected with taxation." By stating the object the informant will be put at ease and thereby will give, more accurate data when

solicited with mandatory powers, or more information if voluntary.

3. The questions and the general character of the schedule should be drawn up with a full understanding of the type of person to whom it is addressed. In advertising, firms use different newspapers and magazines to attract different types of people; the schedule, too, must be couched in language and terms familiar to the informant. Some forms of technical business terminology would be unfamiliar to the country storekeeper, although he probably uses the same process, calling it by a different name. A rigid definition of the terms will often make the most intricate terms comprehensible. However, if the data are being solicited by mail, the prospective informant may feel that he has not the time to study out the desired answers to the schedule.

4. Schedules should at all times contain definitions of the terms used, so that accuracy of statement may be obtained. In the illustrative schedule referred to the terms "farm operator," "farm tenure," "value of implements," etc. are defined, and therefore the individual knows what such terms include and exclude. If the terms are not rigidly defined there may be little homogeneity in the tabulated data of the schedules. In a personal investigation this would not be true, as the information solicited would conform to but one standard.

5. The questions on the form should not *appear* too inquisitorial, but should be clear, intelligible, and preferably answered by the affirmative, the negative, or by numbers. Long general statements are to be avoided.

6. Questions should not require personal judgments as to habits. Some colleges now require the applicant for admission to fill out a schedule which includes such directions as: "Judge yourself carefully according to the following grades—excellent, good, average, poor, and deficient in the following—cheerfulness, punctuality, orderliness, sociability, honesty, courage, reliability, etc." The value of such answers, it will generally be conceded, is very indeterminate. Applicants desirous of college training will not be prone to register their deficiencies. Such questions may have some value if corroborated by data from other sources but, in any case, the first schedule is of little value.

7. As far as possible, corroborative answers should be included in the same schedule of the data. The length of schedule and

cost of collection, however, determine the efficacy of adding such questions.

8. When schedules include the choice of alternatives the method of indicating choice should be definite, whether the choice is to be shown by a check mark or the alternative crossed out. In the schedule referred to it will be noticed that the enumerator is instructed to write "yes" or "no" and not merely check the attribute required. Such uniformity of practice facilitates classification in the central office.

9. The schedule should not require that mathematical computation of data, such as percentages of total, etc., be done in the field by the enumerator. Enumeration is counting, not analyzing the figures counted; analyzing can best be done in the office. But to facilitate office work and afford a brief résumé of the contents of the schedule, spaces should be left on the form for a synopsis of the data. This is shown on the example by the spaces marked: "This space for office use only." In the private schedules already referred to, the student will find the informant so instructed: "Do not write in these columns."

10. The result of an inquiry should justify the expense and time in collection. Unnecessary questions or those for which time cannot be given for analysis should not be included. As an illustration of the cost of one question the census is again quoted:

> In the law for 1920 the words "domestic animals" were omitted and the words "live stock" substituted. To one not acquainted with the technique of census work this appeared to be a perfectly harmless change, but it was presumably made on the supposition that an inquiry in regard to poultry raised in cities would or could be added to the schedule to be applied to barns and enclosures not on farms. Such an inquiry would increase the cost of the field work by about $50,000 and add a proportional amount to the cost of the office work; in addition, it would increase the burden of the enumerators.[1]

There is a great temptation, in drawing up a schedule, to include many questions upon which material would be welcome but cost, time, and results should be the criteria in selection.

11. From the viewpoint of the informant, simplicity, directness, and ease should be the objects sought for; this is particularly true when the information is obtained by correspondence.

[1] Steuart, William, "The Conduct of the Fourteenth Census," *Amer. Stat. Pub.*, p. 573, March, 1921.

Accuracy, comprehensiveness, and yet economy of time and money should dominate the investigation.

Reverting now to the building analogy, the schedule may be likened to the forms used in the making of the concrete foundation. The cement and the material are poured into the forms and take the proper dimensions and design; weak forms yield weak or misshapen foundations. The symmetry and the strength of a statistical structure are also dependent, to a considerable extent, upon the type and character of the schedule. *The contents and the form of schedules, as well as the persons solicited, should, therefore, be fully considered.*

18. Instructions in the Gathering of Material.—The forms used in the construction of a building foundation differ in size and design and, therefore, have to be placed in the right position. The workman, by means of the symbols on the forms, recognizes their purpose and place. The schedule used in statistics may not permit the inclusion of full instructions; the symbols given may require a workman for their interpretation. This interpreter is the enumerator.

When data are collected by enumeration, the enumerator puts all the questions to the informant and records all the answers on the schedule. He, therefore, must have a complete understanding of the form. Because the enumerator is the interpreter, the schedule may be made more complex than one used for an investigation conducted by correspondence, this places a great responsibility upon the enumerator. In addition to class instruction, government enumerators are given a book on instructions containing the duties, powers, and conditions of the enumeration. From this book, the student may form some idea of the task of an enumerator, and understand the rigidity of the definition of the terms used. The enumerator's work is not easy and the whole value of the census depends upon the reliability of the data recorded. It seems to be the consensus of opinion that the enumerators employed by the United States government in censal years should be given more instruction and greater remuneration so that the data may have a greater reliability. *Any problem requiring the collection of data by enumerators demands that all enumerators be given adequate instruction in the definition of terms, and in the time of collection, and have a thorough understanding of the territory assigned, together with the persons or firms to be canvassed.*

In sending out schedules, by mail or through correspondents, the time within which a reply is desired should be definitely stated, so that the office staff may be organized and ready for the classification, tabulation, and analysis of the data. As one writer has put it, "the data must be subjected to the right development, neither being underdeveloped through too hurried work nor overdeveloped through needless device and methods." Time limits upon the reporting of the data make possible the allowance for proper development of the data. The legal time requirements for the publication of the United States Census Reports have been criticized because the time element causes much useful material to be buried in the vaults at Washington.

19. Summary.—This chapter may be summarized briefly by stating that primary data are analogous to the raw materials used in the foundation of a building; they should be well chosen, and at all times representative of actual conditions. They should not be collected at an expense which makes the building of an adequate superstructure impossible. The material must be homogeneous as to the required conditions, homogeneity being obtained by recording on a schedule the information solicited by enumerators, correspondents, informants themselves, or by personal investigation. The schedule must be so constructed as to inspire confidence in the informant, secure accuracy of the data recorded, and amply repay the time and expense of collection and analysis. The schedule is the form which molds the foundation and, therefore, must be constructed according to a well-developed plan which gives full consideration to the nature and problems of the structure to be reared.

The materials entering into a statistical problem may consist of primary data, the utility of which depends upon the aims and purpose of the study. All foundation walls must be able to stand the stress and strain of investigation, otherwise the time and energy may be wasted.

With the assembling of the raw material in the central office, the problem is now ready for the next process, that of classification and tabulation.

References

Bowley, A. L.: "An Elementary Manual of Statistics," chap. VIII.
——: "Elements of Statistics," chap. III.

HILL, J. A.: "The Essentials of a Good Census," *Quart. Pub. Amer. Stat. Assoc.*, March, 1920.
KING, W. I.: "Elements of Statistical Method," chap. VI.
SECRIST, H.: "An Introduction to Statistical Methods."
——"Readings and Problems in Statistical Methods," chaps. II and VI.
STEUART, WILLIAM: "The Conduct of the Fourteenth Census," *Quart. Pub. Amer. Stat. Assoc.*, March, 1921.

Problems

1. If you were investigating the labor turnover in a particular industry for the entire United States, what criteria would determine your method in collecting the primary data?

2. Would you follow the same procedure as in problem 1 if the investigation were confined solely to your own city? Explain.

3. What are the various methods used in the collection of primary data?

4. What are the special problems connected with each method used in the collection of data?

5. What are the deciding and limiting factors in drawing up a schedule?

6. (*a*) Draw up a schedule for recording the Quarterly Unemployment, its Causes and its Relation to Union Membership for All Industries in State X.

(*b*) How would you obtain the data, and who would be your informants?

7. If information were solicited through correspondence, would the schedule be identical in contents and form with that used for the mandatory enumeration process? Explain.

8. What are some of the requisites of a good schedule? Illustrate your answer by applying it to a problem which aims to measure age, occupation, and income of the people in district X.

9. What should be the duties of an enumerator?

CHAPTER V

CLASSIFICATION AND TABULATION

I. INTRODUCTION

Statistical methods include all those devices of analysis and synthesis by means of which statistics are scientificially collected, and used to explain or describe phenomena, either in their individual or related capacities.[1]

The initial steps in any statistical study, as already indicated in the previous chapters, are:

1. To have the problem definitely in mind.
2. To define rigidly the unit of measurement.
3. To collect the data with reference to the unit of measurement.

To pursue these three steps results in the compilation of statistics, which in themselves mean nothing. The data must be treated in a methodical way in order to develop the information that lies buried within the mass of figures. This suggests, at once, the need for some method of analysis.

II. CLASSIFICATION

20. Definition.—The first step towards analysis must be one that will rearrange the mass of data into a definite order, according to a predetermined plan. This step is called "classification." It involves the breaking up of the data already having one common characteristic into other homogeneous groups, each having a characteristic peculiar to itself. As an example, the problem might be to find the mean, median, and modal age of all students in a university, and to classify them according to schools, *i.e.*, law school, school of medicine, etc. All these data have the common attribute of age but the problem involves not only different ages, but also different ages in different schools. The study suggests, first, breaking up the data into age groups

[1] SECRIST, H.: "An Introduction to Statistical Methods."

and then reclassifying according to school. The data, therefore, are sorted or classified twice.

21. Mechanical Aids in Classifying Data.—Classification, therefore, involves sorting the data of the schedules in an orderly way—making order out of chaos. Orderly recording of data can rarely be done by direct transcription. It is customary in offices where hand methods are used to transcribe the material on cards; so that all the detail may be given in a much smaller compass than the schedule, and, furthermore, by the use of code or symbols a small card may be made to contain a wealth of information (see Chart 2). With definite spaces allocated to specific facts, these cards may readily be sorted for any given characteristic. This entails, of course, resorting for every attribute. As a check upon sorting, it is usual to *punch* the possession of a given characteristic in the space allotted to it. The illustration below, while a simple one, illustrates the saving of space and the ease in sorting.

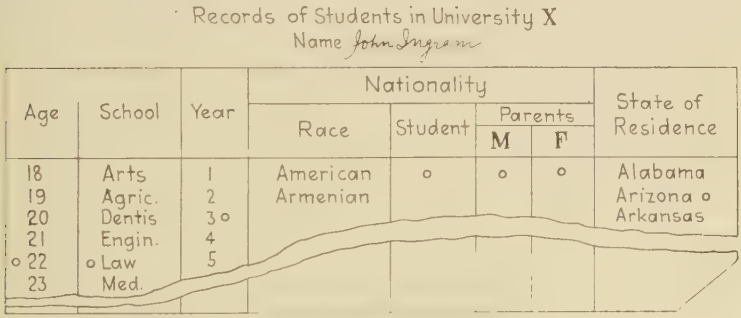


CHART 1.

This small card contains considerable information, and could be made to contain much more if symbols were used instead of actual words. The names of the schools have been abbreviated to save space. A code, however, could be used for all material, and the data for 48 states could then be contained within the area of 48 punched holes. This card, however, illustrates the principle. From this card it is learned that the student is twenty-two years of age, of American birth and parentage, a resident of Arizona, and a third-year law student. If it were the purpose of one of the classifications to find out the number of third-year law students of twenty-two years of age; all cards would be sorted for these attributes and collected in one pile.

It becomes evident that the holes punched for these attributes afford an excellent check. The card, it will be noticed, has the right-hand corner cut off, Thus providing a way of indicating that the cards in the pile are right side up. Since all cards are identical in printing, the holes in the stack of cards for any one attribute should be in direct alignment, so that a ray of light admitted through the pile indicates correct sorting. If one hole is blocked, errors have been made in the sorting. Such a method affords a good check against including irrelevant material. The cards so tested may now be counted and the record made of the number of third-year law students of the age of twenty-two. This process is continued for every desired classification.

In large offices the sorting is done by electrical machines. The cards, drawn up in code, are punched according to the data by a punching machine, which operates very much like a typewriter. The cards are then placed in the sorter, and, to insure that none are upside down, one corner of the card has been cut off. Any irregularities are at once evident, because of the projection at this corner (see illustration below).

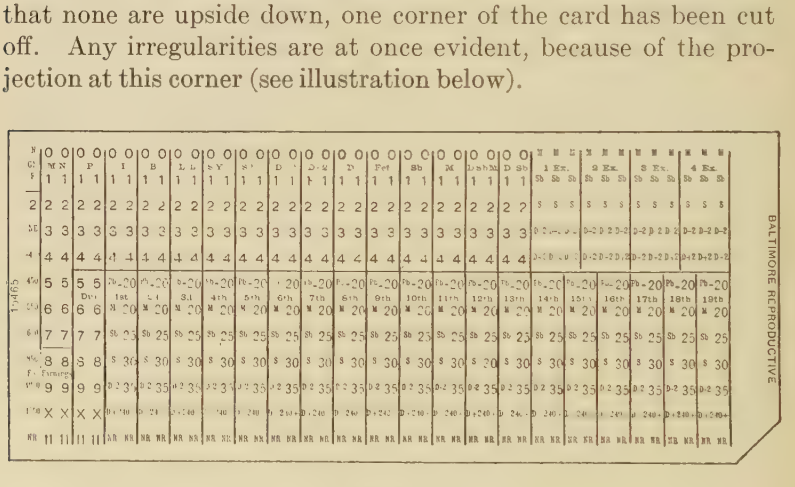


CHART 2.

These cards are now placed in the electric machine, where, by a process of electric needles, the cards are so sorted that contact is made through the punched holes, thus bringing all cards having like characteristics into alignment. From this machine the cards are taken to the electric tabulator, which by means of electric contacts registers the occurrence or count of the various attributes.

22. Care in Editing Data.—Occasionally, the data require consideration before they can be thrown into tabular form, as the editor in the classification process is not merely a sorting machine. In personal investigations this problem rarely arises because of the personal supervision given, however it arises frequently with data collected by agents or through correspondence.

Data may be wrongly recorded on the schedule during the collection process and the editor in classification may discover items which seem abnormal. These abnormalities may be real, or they may be inaccuracies in the recording of the data; therefore, *the editor or classifier of data should never reject or change material without verification for rejection or change.* Such verification gives accuracy to the series. Any apparent abnormality may sometimes be checked through corroborative data contained on the schedule or by referring it to the source of the information. Bias must not enter into any compilation; therefore the editor's judgment of an abnormality should be verified before rejection, or before any changes are made in the recorded data.

Another problem closely related to the above, inaccuracy of statement, frequently arises in the use of collected data. Data, like a chain, are as strong as the weakest link; if some of the numerical information is recorded to the last digit, while other schedules listing the same material have recorded it to the nearest hundred, the editor must give consistency to the tabulation of the data by recording them in round numbers. If the last digits were included in some and not in others, the total of such data would naturally be carried out to the last digit, thus giving a false impression of accuracy.

Another problem of the same character is that of using too general a statement. To some enumerators an engineer may mean a civil engineer, to others a stationary engineer; an artist, to some, may be a general term, to others applying only to one field, that of painting. The editor must have uniformity in the use of terms and by reference to other answers on the schedules it may be possible to make the designations consistent with his classification. There may be inconsistencies within the data recorded on a schedule—age may be too young to indicate married state and yet the record read as "married;" birthplace may preclude being American-born; wages belittle the unem-

ployment recorded; profits be disproportionate to the business recorded, etc.—all these may appear to be inconsistencies and are largely due, *if inconsistent*, to negligence on the part of the agent securing the information, or to indifference of the individual or firm solicited. It is the duty of the editor to try to smooth out these inconsistencies.

In classification, therefore, the editor is not a mere sorting machine, but rather checks for accuracy, uniformity, and consistency of data, the check not being his personal opinion but verification through corroborative evidence contained within the schedule or solicited from the original source of the information. The task of editing tends to become greater when the enumeration process is adopted because the materials are collected by many agents, the scope for error thus being widened.

23. Summary.—Classification is not always a simple thing, involving the presence or absence of a given attribute; it may also demand knowledge of the relative degree to which an attribute is possessed. This involves the problem of class intervals and limits as applied to frequency tables, a subject which will be taken up in Part II. For the present the necessity of knowing in a general way what should be included in, or excluded from, the process of classification is sufficient.

The segregation of facts is but one step towards analysis. The object of any study is to present these facts in an orderly manner, and if the data have been compiled in an endeavor to show certain relationships (derived tables) the data must be recorded in such a way as to emphasize these associations. This requires tabulation, which is the placing of the classified data into tables—the making of a series. A series is "the succession of facts having some common characteristic."[1]

III. TABULATION

24. Kinds of Tables.—A table is a flat surface of two dimensions, length and breadth, the data being arranged in such a manner that they may be read, not necessarily as a line of text from left to right but also up and down, and in some cases diagonally across the table.

Tabulation aids in the visualization of facts, it saves time and space, and aids in analysis. Tables may be divided into two groups:

[1] KELLEY, TRUMAN L.: "Statistical Method," p. 2.

1. General-purpose or primary tables.
2. Special-purpose or secondary tables.

A general-purpose table is a repository of facts gathered not for the purpose of showing *definite* relations, but compiled as a record of facts on some general topic for a particular time or over a period of time. Many such tables are issued by the Bureau of Census. As such tables are the repository of facts, they must, of necessity, record the data to the last digit, *i.e.*, the population of the United States, according to the 1920 Census, was 105,710,620. These figures give pretension of accuracy which does not exist, but the table, to be primary, must be an accurate record of every normal item collected, showing neither bias nor questionable facts in the tabulation.

It must not be assumed, however, that primary tables have nothing to do with analysis. Every table must be constructed according to a definite purpose, and, therefore, possess analysis, but it is the "analysis essential to effective enumeration and tabulation."[1] Table I illustrates a primary table.

TABLE I.—POPULATION OF THE UNITED STATES BY GEOGRAPHIC DIVISIONS, YEAR 1920

Source: Bureau of Census

Territory or division	Population
Total United States	105,710,620
New England	7,400,909
Middle Atlantic	22,261,144
East North Central	21,475,543
West North Central	12,544,249
South Atlantic	13,990,272
East South Central	8,893,307
West South Central	10,242,224
Mountain	3,336,101
Pacific	5,566,871

A special-purpose or secondary table is, as the name implies, a table constructed for a special purpose. Such tables are usually

[1] DAY, E. E.: "Standardization of the Construction of Statistical Table," *Amer. Stat. Pub.*, p. 61, March, 1920.

TABLE II.—AN ESTIMATE OF THE NET SAVINGS OF THE UNITED STATES AND ITS PEOPLE

A	B	C	D	E	F	G	H
Year	Total national income measured in 1913 dollars (millions)	Total national income measured in dollars of given year[1] (millions)	Index of saving in the United States	Ratio of income saved to index of saving[2]	Total national savings measured in dollars of current year (millions) (D × E)	Fraction of income saved F / C	Total income saved measured in dollars of 1913 (millions) (B × G)
1909	$ 30,608	$ 29,259	$ 4,707	1.0624	$ 5,001	0.1709	$ 5,231
1910	32,176	31,471	5,071	1.0624	5,387	0.1712	5,509
1911	31,397	30,904	3,973	1.0624	4,221	0.1366	4,289
1912	33,985	33,807	4,962	1.0624	5,272	0.1559	5,298
1913	35,436	35,436	4,592	1.0624	4,879	0.1377	4,880
1914	33,067	33,401	3,937	1.0624	4,183	0.1252	4,140
1915	35,150	35,920	7,099	1.0624	7,542	0.2100	7,381
1916	41,669	45,861	11,810	1.0624	12,547	0.2736	11,401
1917	42,377	54,664	8,694	1.0624	9,237	0.1690	7,162
1918	38,830	59,930	−1,736	1.0624	−1,844	−0.0308	−1,196
Total 1910–1918	$324,095	$361,394	$48,402	Period	1909–18	0.1561[3]	$54,095

$$\frac{\text{Savings for 1910–1918 at 1913 prices}}{\text{Income for 1910–1918 at 1913 prices}} = \frac{\$46,115}{\$324,095} = 0.14229 = \text{fraction of income saved.}$$

$0.14229 \times \$361,394 = \$51,423 =$ total income saved, in millions of current dollars.

$$\frac{51,423}{48,402} = 1.0624 = \text{ratio of income saved to saving index.}$$

By permission of W. I. King.

[1] Revision, according to information at hand on June 25, 1922, of estimates presented in "Income in the United States."
[2] See computation at foot of this table.
[3] $\frac{\text{Sum of F}}{\text{Sum of C}}$

compiled from the data contained in primary tables and are arranged to emphasize the analysis of the results of a special study. In such tables the data emphasized are thrown into bold relief in order that the special purpose may be indicated and the analysis emphasized. Table II illustrates these points.

A secondary table has been called a "summary text or an analysis table." Because it is a summary table and a derived one, the figures may be given in round numbers and not expressed to the last digit as in primary tables. Such practice, however, obligates the tabulator to indicate on the derived table the sources from which it was derived.

With an understanding of the two types of tables it can readily be seen that tabulation requires careful study, for "primary tables show the results of tabulation and the derived table the results of analysis."[1]

25. Types of Series.—Before taking up the methods of tabular presentation it is necessary to recognize certain types of classified data or series:

1. Temporal.
2. Spatial.
3. Component.
4. Frequency.

A temporal series, as the name implies, is one recording and tabulating a succession of facts over a definite period of time. Table III illustrates such a series.

[1] *Ibid.*

TABLE III.—IMMIGRATION TO THE UNITED STATES: 1899 TO 1923, BY RACES

General Race Classes

(For years earlier than 1899 figures are not available)

Source: Monthly Labor Review, January, 1924

Year	Northern and western Europe	Eastern and southern Europe (including Hebrew)	Southern American	Oriental	Other races	All races
1899	100,187	203,890	1,791	5,070	777	311,715
1900	116,454	314,071	3,114	13,958	975	448,572
1901	128,924	348,100	2,330	7,768	796	487,918
1902	162,840	464,795	3,771	16,198	1,139	648,743
1903	236,105	589,708	5,905	22,880	2,448	857,046
1904	247,066	532,969	8,866	20,874	3,095	812,870
1905	288,295	705,475	10,692	18,066	3,971	1,026,499
1906	269,936	801,054	8,793	16,126	4,826	1,100,735
1907	281,322	956,019	8,007	32,705	7,296	1,285,349
1908	233,235	512,882	11,178	19,417	6,158	782,870
1909	209,418	510,168	20,885	5,464	5,851	751,786
1910	276,272	727,431	23,141	6,369	8,357	1,041,570
1911	269,701	567,431	24,992	6,407	10,056	878,587
1912	232,404	559,738	27,630	7,978	10,422	838,172
1913	271,419	889,627	16,587	10,576	9,683	1,191,892
1914	253,855	921,160	19,568	11,619	12,278	1,218,480
1915	142,168	148,798	16,885	11,306	7,543	326,700
1916	127,990	128,214	23,469	11,184	7,969	298,826
1917	122,927	127,545	23,822	11,031	10,078	295,403
1918	42,892	27,991	21,744	11,954	6,037	110,618
1919	71,202	17,628	34,328	11,898	6,076	141,132
1920	165,871	184,903	58,032	11,659	9,536	430,001
1921	206,995	537,144	36,004	11,962	13,123	805,228
1922	129,434	141,621	21,366	11,137	5,998	309,556
1923	274,507	162,695	67,513	9,986	8,218	522,919
Total	4,863,419	11,081,057	500,413	323,592	162,706	16,929,107

A spatial series is one in which the data are tabulated with reference to space, *i.e.*, the comparison of contiguous territories. Geography governs, and in a spatial series for the entire United States the Bureau of Census has made the geographic divisions as shown in Table IV.

TABLE IV.—POPULATION OF THE GEOGRAPHIC DIVISIONS OF THE UNITED STATES, YEAR 1920

Source: Bureau of Census

Geographic division	Population
New England	7,400,909
Middle Atlantic	22,261,145
East North Central	21,475,543
West North Central	12,544,249
South Atlantic	13,990,272
East South Central	8,893,307
West South Central	10,242,224
Mountain	3,336,101
Pacific	5,566,871

A component series shows the relationship of the various parts to the whole. The previous illustration would be a component series if the population of all divisions were summated, thus giving the total population for the United States, which is placed above the other divisions as in Table I.

A frequency series is one in which variable phenomena are measured at one particular time, such as the problem given in this chapter, that of the age by schools of the students in university X. This type of series will be discussed in detail in Part II, which is devoted to the problems related to frequency series. Table V illustrates such a series.

TABLE V.—NUMBER OF DEATHS IN REGISTRATION AREA BY AGE AND SEX, YEAR 1920

Source: Bureau of Census

Age of decedent	Deaths from all causes,[1] registration area:[2] 1920					
	Number			Distribution per 1,000		
	Total	Male	Female	Total	Male	Female
All ages	1,142,558	598,994	543,564	1,000.0	1,000.0	1,000.0
Under 1 year	174,710	99,239	75,471	152.9	165.7	138.8
1 year	36,986	19,698	17,288	32.4	32.9	31.8
2 years	16,892	8,921	7,971	14.8	14.9	14.7
3 years	11,369	5,815	5,554	10.0	9.7	10.2
4 years	8,475	4,401	4,074	7.4	7.3	7.5
Under 5 years	248,432	138,074	110,358	217.4	230.5	203.0
5 to 9 years	27,051	14,461	12,590	23.7	24.1	23.2
10 to 14 years	19,450	10,413	9,037	17.0	17.4	16.6
15 to 19 years	31,259	15,910	15,349	27.4	26.6	28.2
20 to 24 years	43,892	20,296	23,596	38.4	33.9	43.4
25 to 29 years	49,753	23,024	26,729	43.5	38.4	49.2
30 to 34 years	50,050	24,678	25,372	43.8	41.2	46.7
35 to 39 years	52,093	27,381	24,712	45.6	45.7	45.5
40 to 44 years	47,609	25,477	22,132	41.7	42.5	40.7
45 to 49 years	51,959	28,569	23,390	45.5	47.7	43.0
50 to 54 years	58,072	32,071	26,001	50.8	53.5	47.8
55 to 59 years	61,678	33,869	27,809	54.0	56.5	51.2
60 to 64 years	72,780	39,745	33,035	63.7	66.4	60.8
65 to 69 years	76,564	41,060	35,504	67.0	68.5	65.3
70 to 74 years	79,584	41,466	38,118	69.7	69.2	70.1
75 to 79 years	73,995	37,238	36,757	64.8	62.2	67.6
80 to 84 years	53,474	25,362	28,112	46.8	42.3	51.7
85 to 89 years	29,403	13,321	16,082	25.7	22.2	29.6
90 to 94 years	10,376	4,284	6,092	9.1	7.2	11.2
95 to 99 years	2,459	954	1,505	2.2	1.6	2.8
100 years and over	961	351	610	0.8	0.6	1.1
Unknown	1,664	990	674	1.5	1.7	1.2

[1] Exclusive of stillbirths. [2] Exclusive of Hawaii.

26. The Methods of Tabulation.—*a. Title.*—Every table should have a title, which should be short and simple but clear and comprehensive. It should be so worded as to portray the object of tabulation. For a special-purpose table the title should emphasize the analyses, according to their importance, and give precedence according to purpose.

TABLE VI.—DEATHS, BY COLOR, NATIVITY, PARENT NATIVITY, AND SEX, OF DECEDENT

Color, nativity, and parent nativity	Deaths from all causes[1] registration area:[2] 1920					
	Number			Distribution per 1,000		
	Total	Male	Female	Total	Male	Female
Aggregate	1,142,558	598,994	543,564	1,000.0	1,000.0	1,000.0
White	1,007,117	530,173	476,944	881.5	885.1	877.4
Native	769,822	398,446	371,376	673.8	665.2	683.2
Both parents native	454,243	232,194	222,049	397.6	387.6	408.5
One or both parents foreign-born	222,705	115,517	107,188	194.9	192.9	197.2
Parentage unknown	92,874	50,735	42,139	81.3	84.7	77.5
Foreign-born	223,913	122,204	101,709	196.0	204.0	187.1
Unknown	13,382	9,523	3,859	11.7	15.9	7.1
Colored	135,441	68,821	66,620	118.5	114.9	122.6
Negro	130,147	65,426	64,721	113.9	109.2	119.1
Indian	2,300	1,132	1,168	2.0	1.9	2.1
Chinese	1,348	1,226	122	1.2	2.0	0.2
Japanese	1,511	926	585	1.3	1.5	1.1
Other colored	135	111	24	0.1	0.2	[3]

[1] Exclusive of stillbirths. [3] Less than one-tenth of 1 per 1,000.
[2] Exclusive of Hawaii.

In the construction of Table VI the facts, color, nativity, nativity of parents, and sex, are emphasized in the order in which they are given. If the title of the table were to read "Nativity, Sex, and Color," it would be misleading, because it does not conform to the tabulation, which does not give preeminence to nativity. The order of the title should conform to the tabulation itself. This agreement is easily attained, because the purpose of construction governs the title and the title, therefore, indicates tabular arrangement. A well-developed title may solve many of the problems of tabulation.

b. Stub and Captions.—The first dimension of a table employs an arrangement of the data from left to right, forming rows; the second permits the reading of data up and down, which gives rise to columns. The headings of the columns are called "captions," while those for the rows constitute the "stub."

Captions should be so arranged that the boxes break down in size, according to their importance; if related, the subordinate

heading should be placed below the major caption. The space allotted to subordinate divisions should, furthermore, be comparable to the degree of subordination.

TABLE VII.—GEOGRAPHIC DISTRIBUTION OF NATIVE WHITE POPULATION OF THE UNITED STATES ACCORDING TO PARENTAGE AND SEX, YEAR 1920

Division and state	Total population: 1920	Native white					
		Native parentage		Foreign parentage		Mixed parentage	
		Male	Female	Male	Female	Male	Female
United States							
Geographic divisions:							
New England							
Middle Atlantic							
East North Central							
West North Central							
South Atlantic							
East South Central							
West South Central							
Mountain							
Pacific							

In Table VII the major caption "Native white" is broken down into minor divisions for parentage, which, in turn, make the distribution according to sex. There are three divisions for parentage and, being of equal importance, each is allotted the same space, but less than that given to the major heading. Sex is a subdivision of parentage, the length of each parentage box is then subdivided for male and female. The main caption, therefore, reads down from the general to the more specific.

The first difficulties encountered in tabulating classified data are those relating to the choice of the data to be thrown into the stub or caption, and the arrangement to be followed in these two dimensions. There is no general solution to these problems, the answers depend upon the type of table used. The same procedure would not be followed for a general-purpose table as for a special-purpose series, also the method differs according as the series are frequency, temporal, spatial, or component.

c. What Should be Stub and Caption.—In primary or general-purpose tables such problems are determined, largely, by the dimensions of the page. The page tends to be the limiting factor, and, as length exceeds breath, there are more rows than columns;

hence it is customary to throw into the stub the data having the greatest number of groupings. Sometimes the length of the heading determines the material to use in stub or captions; long headings should be placed, if possible, in the stub.

A special-purpose table rarely affords such latitude in construction as a primary table does. It is a table of analysis, and the special relationships should be brought out distinctly—it is not a mere repository of facts but an analysis of such facts. It would be foolish to assume that the dimensions of the page have no influence on the form of the table; in fact, instead of being the deciding factors, as in primary tables, they are the limiting ones. The decisive factor in determining what should be stub or caption is the purpose of the investigation; a plan which will bring out the necessary relationships.

d. The Arrangement of Stub and Captions.—The primary table, once again, affords considerable scope in arrangement. The arrangement, however, should be such as to facilitate the use of the table for reference.

The stub of any table may be drawn up in one of four ways:

1. Alphabetically.
2. Geographically.
3. Chronologically.
4. By magnitude.

Since a primary table is a reference table, the order adopted should be such as to facilitate easy reference and transcription. Where possible, the alphabetical order should prevail, as in the Bureau of Census tables for the racial population of towns. By such a listing the reader interested in town X can readily find it, whereas if the towns were listed according to magnitude facility of reference would be lost. For spatial series, geography governs, as found in the geographic divisions of the United States see Table I. Magnitude should rarely govern the arrangement of the stub in a primary table, for it implies special purposes or analysis of the data.

For a time series chronology dominates the listing of the stub, the most remote date being placed in the upper left-hand corner of the table, the other years reading down in sequence to the most recent. Sometimes tables are constructed with the most recent date at the top. This violates custom. If it is necessary for emphasis to place the recent date at the top, it is better practice

to list immediately below it the most remote, and then work down to the recent.

In a special-purpose table analysis naturally governs the stub and captions, for it must convey clearly the interpretation. Comparison is usually the purpose; therefore the related parts

TABLE VIII.—DEATHS FROM ALL CAUSES IN REGISTRATION AREA OF UNITED STATES, YEARS 1900–1920

Calendar year	Deaths[1] from all causes in the registration area				
	All ages	Under 1 year		Under 5 years	
		Number	Per cent of total	Number	Per cent of total
	A	B	C	D	E
1920	1,142,558	174,710	15.3	248,432	21.7
1919	1,096,436	161,621	14.7	229,813	21.0
1918	1,471,367	193,855	13.2	306,143	20.8
1917	1,068,932	171,024	16.0	243,708	22.8
1916	1,001,921	164,660	16.4	234,081	23.4
1915	909,155	148,561	16.3	203,223	22.4
1914	898,059	155,075	17.3	214,120	23.8
1913	890,848	159,435	17.9	225,129	25.3
1912	838,251	147,455	17.6	204,639	24.4
1911	839,284	149,322	17.8	209,482	25.0
1910	805,412	154,373	19.2	217,319	27.0
1909	732,538	140,057	19.1	196,534	26.8
1908	691,574	136,432	19.7	189,865	27.5
1907	687,034	131,110	19.1	183,774	26.7
1906	658,105	133,105	20.2	186,978	28.4
1905	545,533	105,553	19.3	147,384	27.0
1904	551,354	102,880	18.7	145,902	26.5
1903	524,415	96,857	18.5	139,940	26.7
1902	508,640	98,575	19.4	143,515	28.2
1901	518,207	97,477	18.8	141,678	27.3
1900	539,939	111,687	20.7	164,137	30.4

[1] Exclusive of stillbirths. [2] Exclusive of Hawaii.

should be drawn closely together. This can best be done through the use of columns. The most important place in any table is the upper left-hand corner; therefore analysis requiring special emphasis should occupy this position, or relatively near to it,

according to the degree of emphasis desired. Totals, if they are important, should be placed at the top which violates accounting practice. This rule applies to both primary and secondary tables. Columns showing the absolute figures with their related percentages should be brought into juxtaposition, as shown in Table VIII, columns B and C.

Alphabetical order of the stub is rarely found in special-purpose tables, there being no significance to the mere sequence of letters. For such tables the general rule is: *Purpose of construction should govern both content and form.*

e. Rulings.—With an understanding of what should be stub and captions and their arrangement, the data may now be tabulated into the items of the stub and the captions, but the different data must be ruled off from one another, in order that perspicuity may be given to the finished table. The headings, or captions, should be ruled off box fashion, the various major headings and totals being ruled off with double lines. Table V illustrates major captions and columns. The absolute number of deaths and the distribution per 1,000 are separated by double lines, likewise, the total or aggregate. Minor captions are boxed in with but single lines. The top of the table should have double rulings; the stub, however, should never be boxed in. Subdivisions of the major headings of the stub should be indented, as in Table VI ("White" being subdivided into "Native" etc.).

f. Aids in Reading.—In tabulation the convenience of the reader should be given consideration. He may be aided in the reading of the table in several ways. A lengthy stub or any stub may be broken up into groups of five lines separated by double spacing see (Table II). This arrangement facilitates reading, especially if the captions are numerous. Another aid, if the captions occupy considerable space laterally, is too repeat the stub on the right-hand side of the table. Sometimes the rows are numbered, while the columns should be lettered or numbered in order to facilitate comparison and textual reference. The columns of Table II are lettered.

Diverse units of measurement may be indicated immediately below the respective subordinate captions, *i.e.*, on the first row. If space permits it is better to indicate them in the captions themselves, as in Table II. If a table is homogeneous, measuring but one type of data, the units may be indicated in the title. In derived tables only significant figures are used. Some digits

may, therefore, be dropped, the figures being given in round numbers. The digits dropped may be indicated in this manner: "000's omitted"—thus when 20 appears in the column it means 20,000 (see Table LXX, page 281).

Tables should be so arranged that, if possible, they read with the text of the book and not at right angles to it. Also, if possible, they should be small enough to prevent folding. Finally, the sources of data and the units of measurement should be clearly stated in the table.

IV. SUMMARY

Classification and tabulation in themselves appear to be very simple, but when the considerations to be applied to various data are remembered it becomes evident that the success of tables, especially secondary tables, depends largely upon the method of construction. The foregoing paragraphs have but touched upon some of the points worthy of consideration and the chapter might well be concluded by a partial summary of the standards given by E. E. Day.[1]

It is generally recognized: (1) that every table should be self-sufficing, containing within itself a clear explanation of the meaning of the terms displayed; (2) that every table should be logically a unit, containing only data which are intimately related with one another; (3) that column- and row-headings should be brief, unambiguous, and self-explanatory, table footnotes being used when necessary to make the headings perfectly clear; (4) that coordinate and subordinate relationships among the column- and row-headings should be shown by variations of boxing in the captions and of indentation in the stub; (5) that varieties of letters, figures, lines, column-widths, and interlinear spacings should be employed to facilitate easy and intelligent use of the table; (6) that columns and rows should be lettered or numbered if cross-reference is desirable; and (7) that sources and units should invariably be indicated.[1]

References

Bailey, W. B.: "Modern Social Conditions," pp. 24–31.
—— and Cummings, J.: "Statistics," chaps. I–IV.
Bowley, A. L.: "Elements of Statistics," chap. IV.
——: "An Elementary Manual of Statistics," chap. VI.
Davies, G. R.: "An Introduction to Economic Statistics," pp. 3–9.

[1] Day, E. E.: "Standardization of the Construction of Statistical Table," *Amer. Stat. Pub.*, p. 61, March, 1920.

DAY, E. E.: "Standardization of the Construction of Statistical Tables," *Quart. Pub. Amer. Stat.*, March, 1920.
FALKNER, R. P.: "Statistical Tabulation and Practice," *Quart. Pub. Amer. Stat. Assoc.*, June, 1916.
KELLEY, TRUMAN L.: "Statistical Method," chap. I.
KING, W. I.: "Elements of Statistical Method," chap. IX.
SECRIST, H.: "An Introduction to Statistical Methods."
——: "Statistics in Business," chap. IV.
WATKINS, G. P.: "Theory of Statistical Tabulation," *Quart. Pub. Amer. Stat. Assoc.*, December, 1915.

Problems

1. It is desired to know the enrolment in university X classified according to age, school, class, religion, and race.

a. How many ways may the data be classified?

b. What means would you adopt to facilitate:

(1) Classification of the data?

(2) Accuracy in tabulation?

2. In the above study if one student were recorded as being fifty-two years of age, how would you treat this apparent abnormality?

3. Assume that data have been supplied on each of the following topics:

a. Distribution of the Population of the United States as Urban and Rural, 1900, 1910, and 1920: By States and Geographic areas.

b. Number of Farm Animals in the United States, 1900 to 1923.

c. Per Cent of Union Organization among Occupations in the Building Trades in the United States, 1910 and 1920.

d. Number and Per Cent of New York Firms Distributing Home Work (Garment Workers) and of the Home Workers Supplied with Work.

For each of the preceding problems answer the following questions:

(1) What data would you place in the captions? Why?

(2) What data would you place in the stub? Why?

(3) What order and devices would you use in arranging the stub? Why?

4. Draw up the tabular form for each of the above problems.

5. What types of series are included in the above?
What are their characteristics?

6. Differentiate a special- and a general-purpose table with regard to contents and construction.

CHAPTER VI

GRAPHIC METHODS

I. INTRODUCTION

The previous chapters have been devoted to the processes by which data are gathered into tabular form, which give rise to the various types of series. A table of figures to the average eye, however, conveys little, the relationships existing within the table being very difficult to see. As an aid in the presentation of data, graphic methods are used. In this text the term "graphic methods" includes all sorts of statistical devices for the graphic presentation of data. In it an indication of the principal features of the methods of presentation will be given. For a detailed account of such devices the reader is referred to W. C. Brinton; "Graphic Methods for Presenting Facts," and K. G. Karsten, "Charts and Graphs."

Bowley has defined graphic methods as follows:

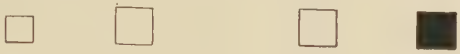
> The main use of diagrams is also (referring to averages) to present large groups of figures so that they shall be intelligible in their entirety, and the test for all diagrams is that the diagram as drawn should afford the best view of the series or group of figures as the eye can appreciate—they are auxiliary rather than essential, merely an aid to the eye and a means of saving time.[1]

II. DIAGRAMS AND MAPS

27. Defects of Area Diagrams.—Graphic presentation is known to all readers. Magazines, newspapers, billboards, and motion pictures, by means of diagrams, interpret figures in a most convincing way. The various devices used, however, do not always, in the terms of the above definition, afford the best view of figures that "the eye can appreciate." The eye is a poor judge of areas, and yet the most popular form of diagrammatic presentation uses areas to indicate relationships. The relative strengths of the navies of the world are represented by different-

[1] Bowley, A. L., "Elements of Statistics," p. 143.

sized ships; the gold reserves in the United States in 1924, when compared to those of 1913, are portrayed by bags of gold varying in size, the purchasing power of the dollar is indicated by means of pie diagrams. Areas require mathematical computation; then why should the eye be given a task greater than the mathematical ability of many people?



Figs. 1 and 2.

To illustrate the difficulties in the visualizing of areas the two squares in Fig. 1 represent a mathematical relationship of 2 to 1. To the average eye it appears to be about 1.5 to 1 because the reader considers the lengths of the sides as criteria of area. They are, but in a square the area varies with the square of the side. Therefore a square having sides of twice the length of

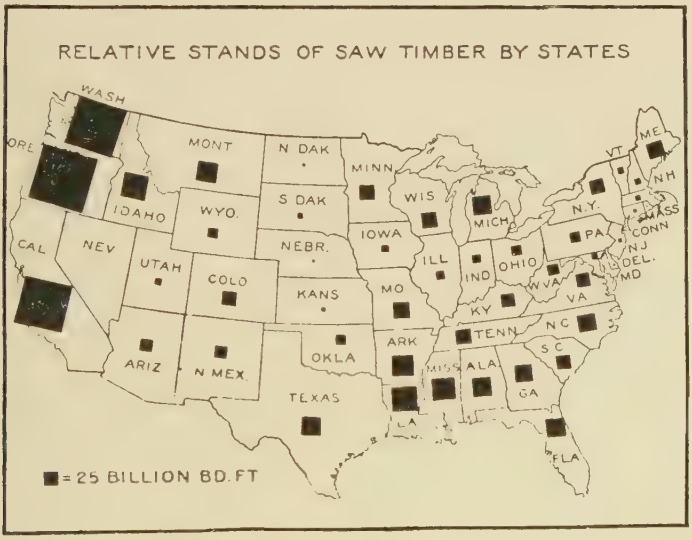


Chart 3.

another would have an area four times as great. In such diagrams, the reader should compare not merely the length of one side of each diagram but the product of two sides of each.

Chart 3 illustrates the difficulties of such methods. Who can approximate, by a cursory glance at the map the number of billion board feet of timber in the state of Oregon?

Colors, too, have an effect on the judgment of areas. Fig. 2 presents two squares of equal areas. The black, however, by absorbing more light appears to be the smaller of the two.

What has been said of squares applies equally well to circles, with the added difficulty of judging the area of the various segments.

From a statistical point of view, and not that of the advertising manager, all devices using more than one dimension should be used with extreme care; the more dimensions used the greater is the confusion in interpretation. Graphic devices should not lie.

28. Secrist's Standards for Graphic Presentation.—Secrist has set up the following standards:

First.—A statistical fact and its form of representation should agree. By this single standard, deception, whether resulting from a confusion of the apparent with the real, or of the superficial with the fundamental, is provided against. The object of business research, as in all research, is the establishment of truth. Standards for graphics provide for their use in influencing, but never in deceiving men. In spite of the standards adhered to, however, both results may be accomplished by the same graphic device. It is their appeal, their smug finality, which suggests their virtues, and at the same time conceals their weaknesses.

Second.—Graphic forms should be selected according to their psychological appeal and their ease of comprehension, care always being taken not to violate the first standard.

Third.—Graphic forms should be chosen in accordance with (*a*) the form and complexity of the subject matter illustrated, and (*b*) the type of consumer for whom they are intended, or the purpose which they are designed to serve.

Fourth.—Graphic devices should be considered more as illustrations of analysis than methods by which analysis is made.

Fifth.—Graphic figures should be drawn as accurately as the visual representation will permit. Accuracy, of course, is never absolute. In graphics, a realization of relative accuracy of each part and the totality is the standard set. To this standard, a corollary is needed; graphic forms should always be accompanied by the original data which they represent.[1]

The above rules, stress the need for care in representation of data by graphic devices, and this point cannot be overemphasized.

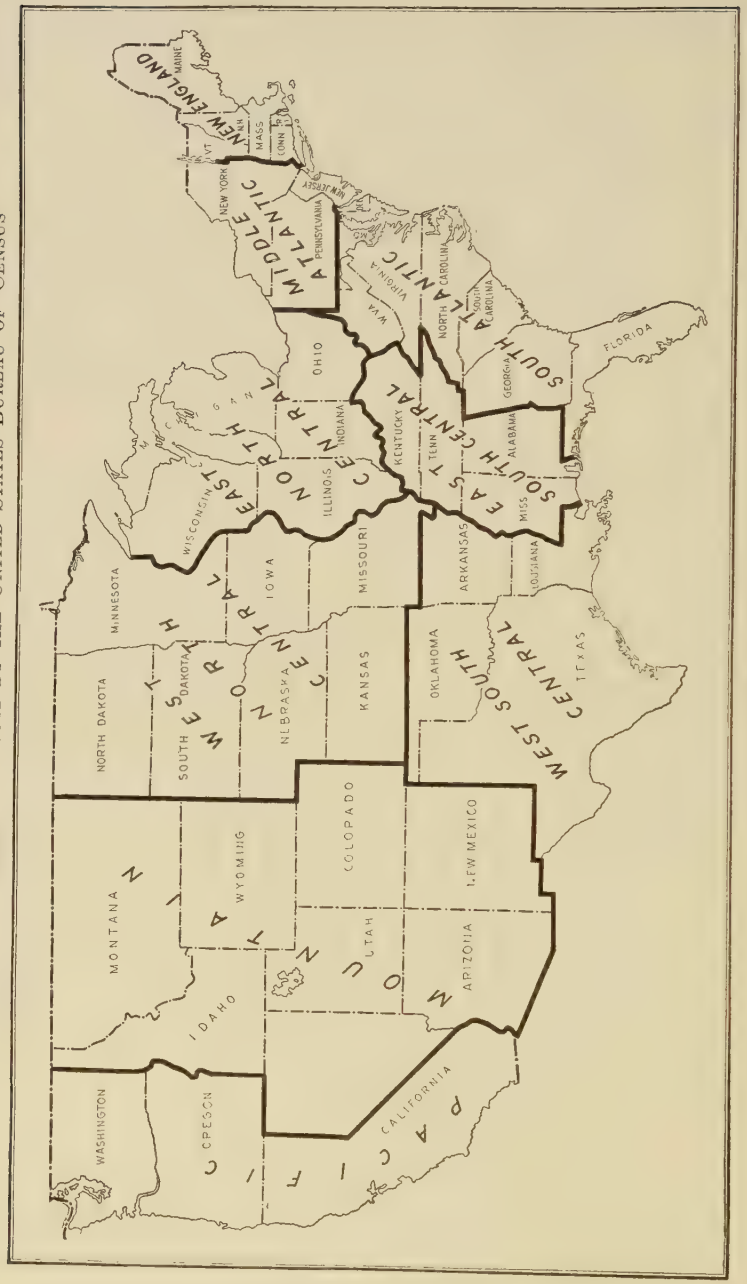
[1] Secrist, H., "Statistical Standards in Business Research," *Amer. Stat. Pub.*, pp. 54–55, March, 1920.

29. Résumé of Methods for Illustrating Different Series.—The various ways of illustrating data may now be considered. Before doing so, however, it may be well to give a very general classification of the types of graphs in use, as they relate to the four types of series. It should be understood that this list is only indicative of type, purpose always dominating choice.

- Component series
  - Pie diagrams
  - 100 per cent bar diagram
  - Bar diagram
- Spatial series
  - Maps
    - Tinted
    - Cross-hatched
    - Dot
    - Pin
  - Pie diagrams
  - Bar diagrams
- Time series
  - Bar diagrams
  - Curves
    - Smoothed
    - Staircase
  - Ratio charts
- Frequency series
  - Curves
    - Cumulative or ogive
    - Lorenz
    - Smoothed
  - Ratio charts
  - Vertical bar diagram

30. Pie Diagrams.—The figures of the population of the United States by geographic divisions lose all significance when read hurriedly down in the tabular form shown in Table IX.

The question arises; how can the significance of these figures be readily realized? The series is a component series. Therefore, if each part is expressed as a percentage of the whole, these figures (column C, Table IX) may be represented on a pie diagram, the various segments of which may be constructed to indicate the relationship of the parts to the whole, as well as the differences between the parts. A difficulty arises in the derivation of segments, that is, how to represent the respective percentages. A circle contains 360 degrees; therefore the figures may be changed into degrees, thus giving the arc through which each segment should pass. This is a somewhat laborious process and also emphasizes the difficulties in appreciating visually the relative values of the segments. Chart 5 illustrates the pie diagram of the data contained in Table IX.



GEOGRAPHIC DIVISIONS ADOPTED BY THE UNITED STATES BUREAU OF CENSUS

CHART 4.

TABLE IX.—POPULATION OF THE UNITED STATES BY GEOGRAPHIC DIVISIONS
YEAR 1920

Source: Bureau of Census

Territory or division	Population	
	Number B	Per cent C
Total United States	105,710,620	100
New England	7,400,909	7
Middle Atlantic	22,261,144	21
East North Central	21,475,543	20.3
West North Central	12,544,249	11.9
South Atlantic	13,990,272	13.2
East South Central	8,893,307	8.9
West South Central	10,242,224	9.7
Mountain	3,336,101	3.2
Pacific	5,566,871	5.3

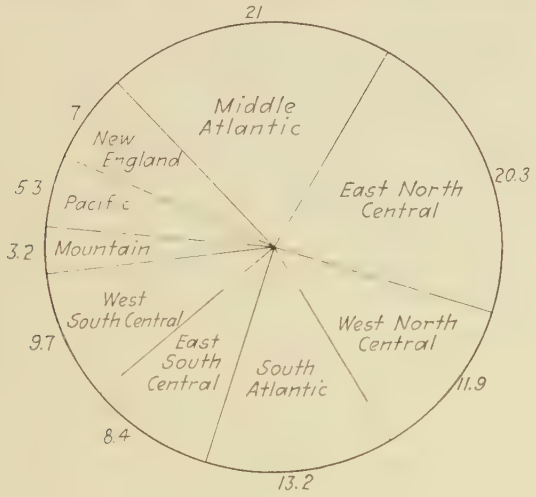


CHART 5.

In drawing pie diagrams the percentage figures, or absolute numbers, should be placed immediately outside the circumference of the circle, and all should be written in the same alignment

in order to permit ease in reading. Furthermore, the wording to differentiate the various parts should be made within the respective segments and in such a manner as to facilitate the reading of all printed matter. If the segments are small, it may be difficult to have all lettering read the same way. Wherever possible, the segments should be so planned as to permit this ease in reading. In Chart 5 *illustration above*, it is evident that, had the New England segment started at the top of the circle, it would have been necessary for the writing to run at right angles to the present percentage figures, making the chart difficult to read. By careful planning this was avoided. This type of chart should be used sparingly.

31. One-hundred Per Cent Bar Diagrams.—If it is necessary to break up a component series into its constituent parts, the

PERCENTAGE DISTRIBUTION OF POPULATION OF THE UNITED STATES BY GEOGRAPHIC DIVISIONS

Source: Bureau of Census

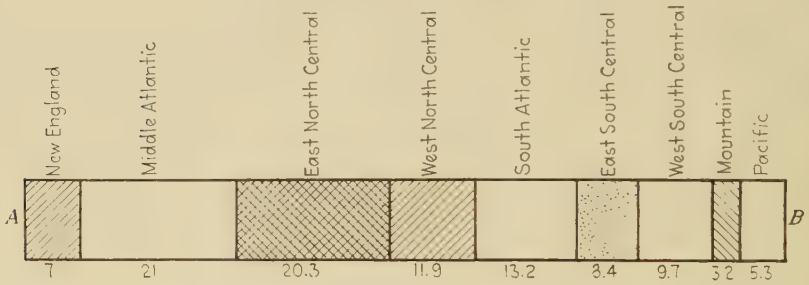


CHART 6.

100 per cent bar diagram gives a much better picture of the relationships and is much easier to construct than the pie diagram Chart 6. The figures in such diagrams are relative and not absolute. Chart 6 presents the same data as Chart 5. In Chart 6 the length of any part of the bar, for any geographic area, represents its percentage of the whole, the whole being represented by the distance between *A* and *B*, which equals 100 per cent. The geographic divisions are differentiated by means of different cross-hatchings. The width of the bar merely affords space for such. The illustration is self-explanatory. Its ease of construction and comprehension is apparent when compared to Chart 5.

Chart 7, a composite bar graph, indicates the scope to which such diagrams may be used.

Before describing the other graphic methods it would be well to emphasize the fact that all charts should indicate the sources from which their data were compiled, and for charts other than 100 per cent diagrams the original data should accompany the diagram, being given preferably within the chart, so as to afford easy reference, or in tabular form near the chart.

COLOR OR RACE, NATIVITY, AND PARENTAGE, BY DIVISIONS
1920, 1910, and 1900
Source: Bureau of Census

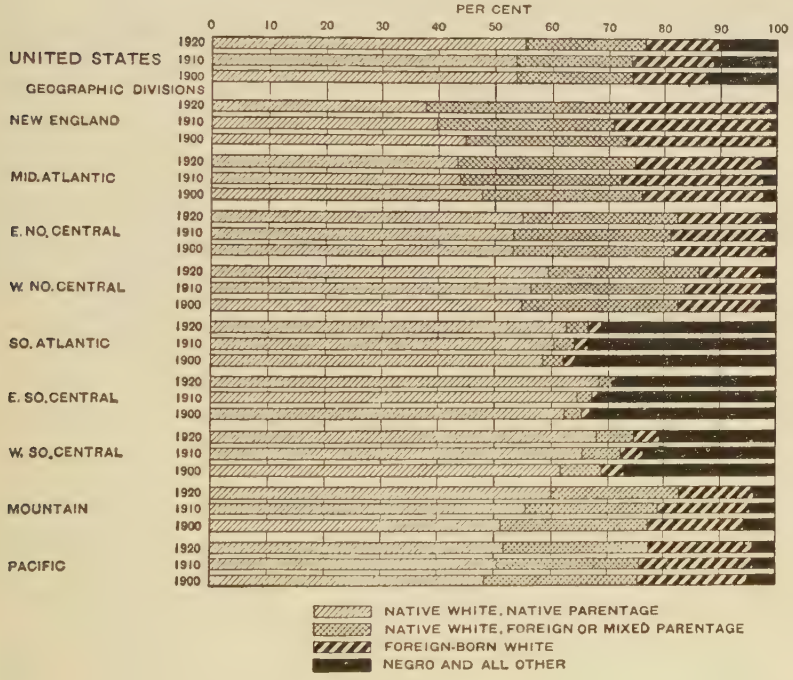


CHART 7.

A chart is as weak as its own data and a chart maker must carefully weigh and consider his data before permitting himself the pleasure of illustrating them with a chart.[1]

When the data are attached to the chart the reader may judge whether or not the chart maker was justified in such presentation; without the data, no check is possible.

32. Bar Diagrams.—It is not always desirable to express numbers as relatives, that is, in percentage figures. The purpose

[1] KARSTEN, K. G., "Charts and Graphs," p. 48.

may be to show relations and yet preserve the absolute figures. For such problems the bar diagram may be used. Naturally, equal distances on the bar will represent equal changes in magnitude (absolute numbers). The bars are given width to make them perspicuous; all bars must be of equal width. Chart 8 is a bar diagram chart. In it the data are found within the field. The best place for indicating the data is immediately at the left of the bars. The data should never be written at the end of the bars, as this spoils the perspective—the long bars, having more figures than the shorter ones, would tend to have their length relatively extended. Figures written *within* the bars also make it difficult to visualize the relationships which exist between the

VALUE OF PETROLEUM AND NATURAL GAS PRODUCED IN UNITED STATES, YEAR 1920

Source: Bureau of Census

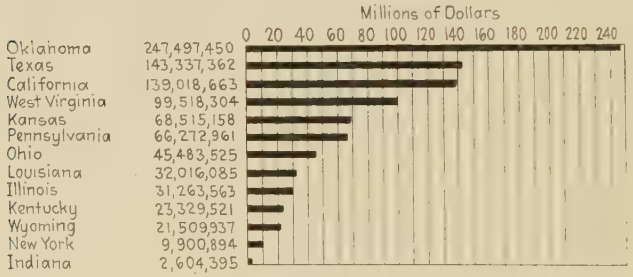


CHART 8.

bars. Whenever possible, the data should be placed at the left of the bars or in a separate table near the chart. The arrangement of the stub for such graphic presentation follows but one rule, namely, purpose; magnitude governs in this illustration.

a. Need for Horizontal Zero.—In bar graphs the zero for the attribute should, when possible, be indicated in the chart. This need will become apparent in the study of Chart 9.

Had the composite bar chart (Chart 9) had continuous unbroken bars (the bars starting at the age of forty), then from the pictorial point of view the expectancy of life for white males would have doubled in the last decade, which, of course, is absurd. In order to prevent such conclusions and yet maintain a large scale, the zero point is indicated from a zero ordinate. Near the zero lines the bars are broken. This break indicates that a part of the bar is missing, about thirty-nine years. The gap at once arrests the attention of the reader and prevents the possi-

LENGTHENING THE LIFE SPAN OF INSURED WAGE EARNERS

GAIN IN EXPECTATION OF LIFE AT AGE 10

WAGE EARNERS INSURED IN METROPOLITAN LIFE INSURANCE COMPANY
1921–1922 AND 1911–1912 COMPARED

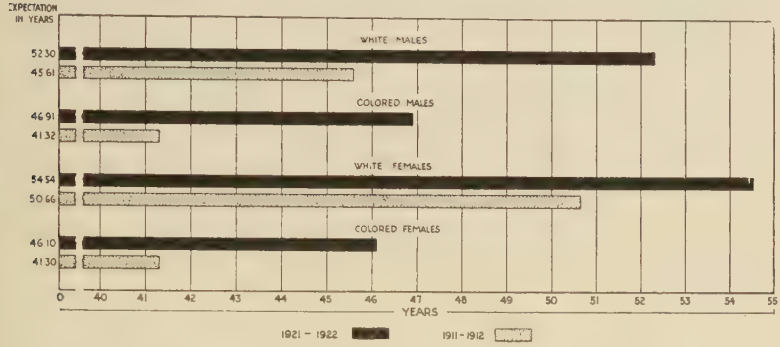


CHART 9.

SOUTHERN BORN NEGROES IN NORTHERN UNITED STATES, 1910 AND 1920

Source: Monthly Labor Review, March, 1924

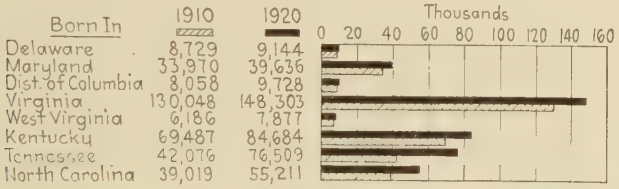
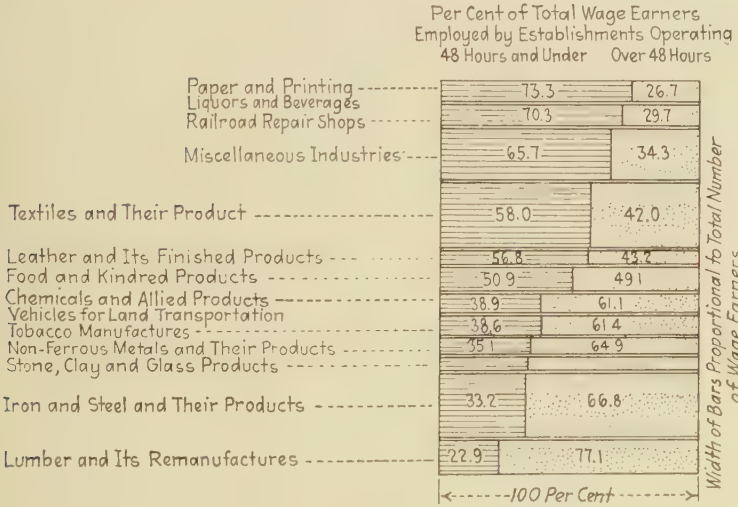


CHART 10.


Per Cent of Total Wage Earners Employed by Establishments Operating
48 Hours and Under
Over 48 Hours
Paper and Printing 73.3 26.7
Liquors and Beverages 70.3 29.7
Railroad Repair Shops
Miscellaneous Industries 65.7 34.3
Textiles and Their Product 58.0 42.0
Leather and Its Finished Products 56.8 43.2
Food and Kindred Products 50.9 49.1
Chemicals and Allied Products 38.9 61.1
Vehicles for Land Transportation
Tobacco Manufactures 38.6 61.4
Non-Ferrous Metals and Their Products 35.1 64.9
Stone, Clay and Glass Products
Iron and Steel and Their Products 33.2 66.8
Lumber and Its Remanufactures 22.9 77.1
100 Per Cent
Width of Bars Proportional to Total Number of Wage Earners

By permission of National Industrial Conference Board.

CHART 11.

bility of such absurd conclusions as intimated above. From this statement it follows that all charts should indicate the zero line of the attribute.

CHANGE IN POPULATION ELEMENTS IN CONNECTICUT BETWEEN 1870 AND 1920

Source: Monthly Labor Review, Jan., 1924

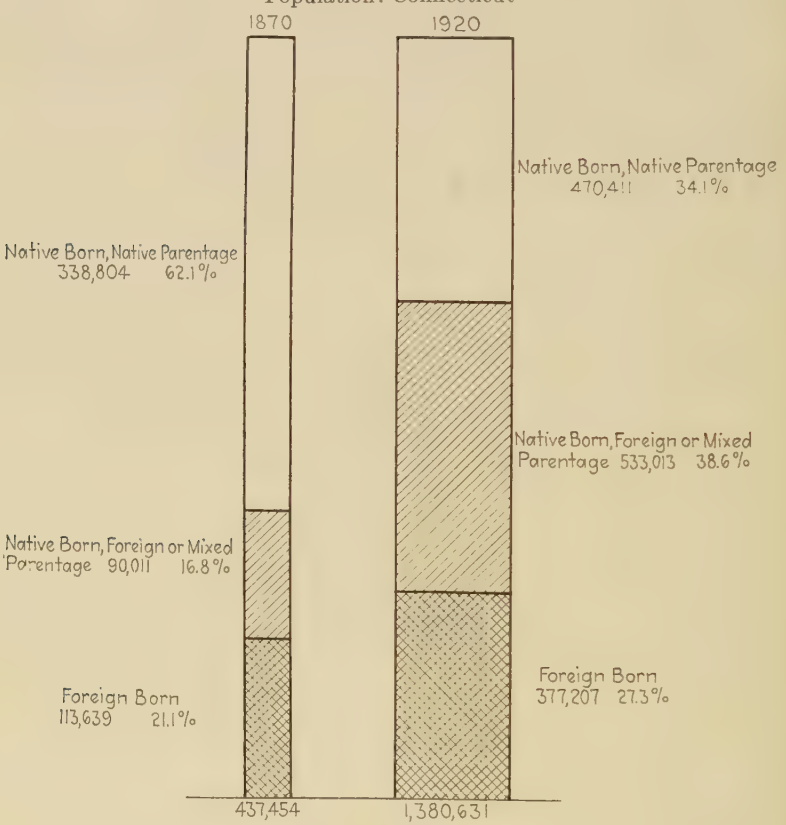


CHART 12.

Chart 10 illustrates a composite bar diagram, the arrangement of the stub being spatial.

Bar diagrams may be made to contain a wealth of information, but it should be remembered that ease of perception is one of the

rules of graphics. Chart 11 illustrates a complex bar graph, it is a composite 100 per cent diagram on all sides. It will be noticed that the figures are contained within the bars, owing to the impossibility of placing them on the exterior. This chart is really too complex to perceive clearly the many relationships.

Another type of chart which is in common use is the vertical bar chart, in which the bars are vertical instead of horizontal. The vertical position makes it rather difficult to label. However, it is a very effective method in illustration. Chart 12 is a 100 per cent vertical graph which is self-explanatory.

Chart 13 illustrates a composite vertical which is often used in portraying certain types of time series. Such a graph is called the staircase curve.

33. Maps.—*a. Types in General.*—Probably the most commonly used device for depicting spatial series is the statistical map. Maps may be divided into three classes;

1. Tinted maps.
2. Cross-hatching and dot maps.
3. Pin maps.

On tinted maps the blending of colors represents certain aspects of the data. Such maps are expensive to print. For illustrations see "A Graphic Analysis of the Census of Manufactures," by the National Industrial Conference Board. The chief problem in the making of tinted maps is the blending of colors in a manner to indicate clearly the changing relationships.

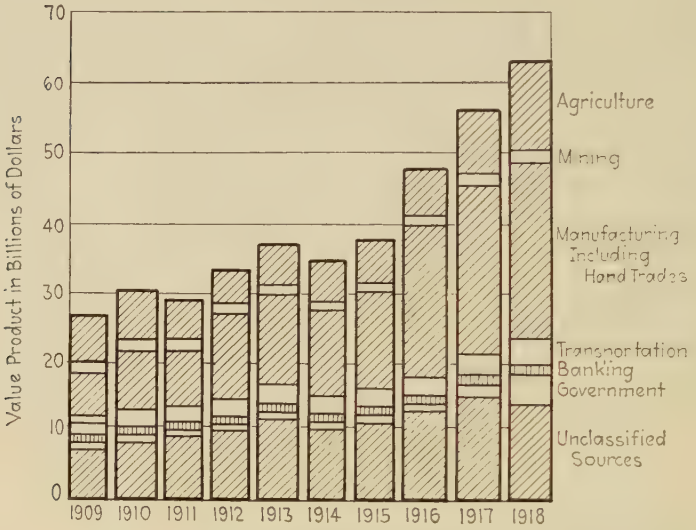
b. Static Maps.—For general work black-and-white maps are generally used. In order that the various attributes of the data may be differentiated, it is necessary to use many devices in indicating these two colors. The method is known as cross-hatching.

(1) Cross-hatched Maps.—Different shadings and designs are used on the same map, dark shadings usually being used for the extreme items above the median group and light for the extremes below. The shading must be well chosen in order that the relationships may be given emphasis. In such maps the shades correspond to the class groupings and not the frequency within any one group. Chart 14 indicates such practice.

(2) Dot Map.—The dot map is similar in design to the cross-hatched type, dots replacing the various hatchings. The dots are very satisfactory if the data are not too concentrated

CONTRIBUTIONS TO THE NATIONAL INCOME MADE BY VARIOUS INDUSTRIES, 1909–1918

Based upon Table 2



By permission of the National Bureau of Economic Research.

CHART 13.

PER CENT URBAN IN TOTAL POPULATION, BY STATES, 1920

Source: Bureau of Census

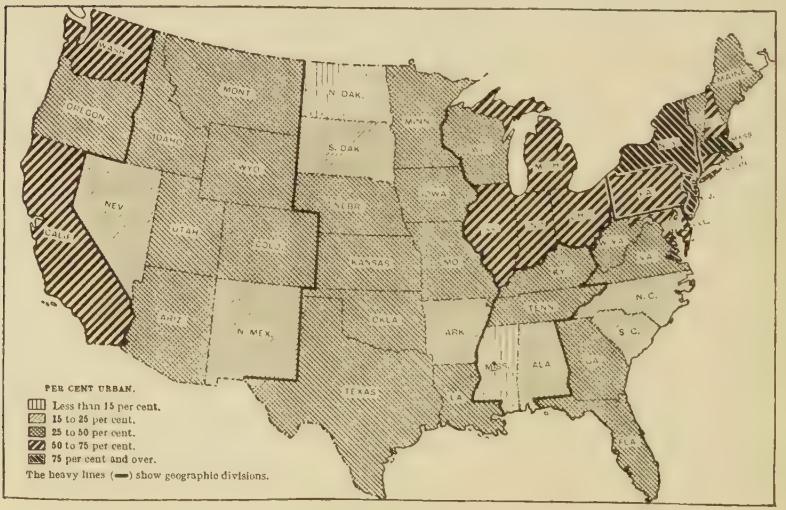

PER CENT URBAN.
Less than 15 per cent.
15 to 25 per cent.
25 to 50 per cent.
50 to 75 per cent.
75 per cent and over.
The heavy lines (—) show geographic divisions.

CHART 14.

in any one place, thus permitting the use of dots of uniform size. The dots, though of uniform size, may be divided into quarters, so as to enable a fivefold classification of data, as indicated in Chart 15. In many cases, however, the data cannot be so concentrated, the range being too great. Resort is then made to different sized dots, a large one being used for large frequencies. Such a method spoils the perspective of the map, the problem of areas vitiating the ease of comprehension. When constructing dot maps dots of uniform size should be used.

ALL CATTLE ON FARMS
Number, by States, April 15, 1910
Source: Bureau of Census

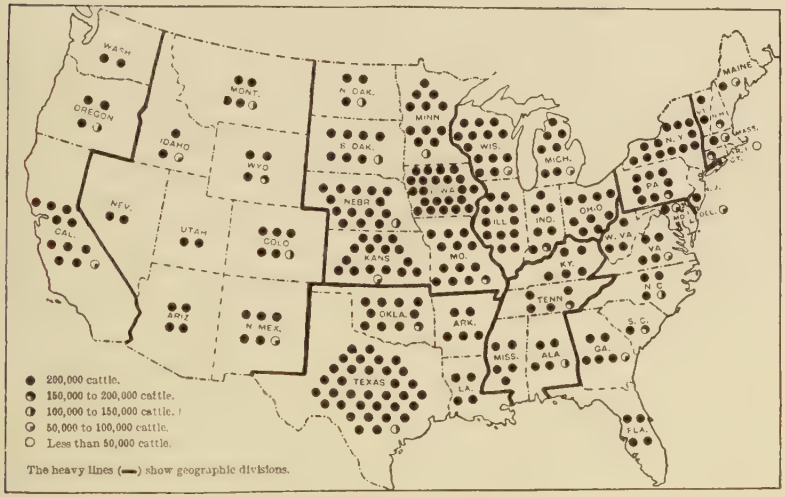


CHART 15.

(3) Limitations of Static Maps. —These statistical maps, it must be apparent, emphasize both position in relation to magnitude and relationship according to magnitude. A chart cannot present very readily all this information because it cannot throw into bold relief the two types of information. However, graphs, at best, are only approximations.

c. Dynamic Maps.—Cross-hatched and dot maps are static maps. They portray very effectively past events, but are unable to be modified with the changing conditions and so new maps have to be made. The business man wants a dynamic map—one that may be modified with time. A graphic study of sales by territories would require a map showing at the end of each

day, or period, the actual conditions in the respective territories clearly. Such a map must be constructed of something that is movable. The pin map permits such changes. The reader is familiar with this type. It was the favorite device used during the Great War to indicate the change in the battle front. The pins or flags which represented the contending armies were moved with every change in front. For the business man a pin map is extremely useful; a pin designating the unit x or number of units of sales, etc., is placed according to current data. In such a map it is desirable to use pins with different-sized heads, in order that the busy territory may be emphasized in the picture. By using specially adapted pins a wall map may be constructed.

III. CURVES

34. Introduction.—The curve is the most frequent graphic device used for portraying data. It has not the universal application, however, of the bar diagram, which may be used for practically all types of series, for the curve "can only be used for data of which the stubs form values or readings of a mathematician's variable."[1]

35. Determination of Scales.—Curves require data of two dimensions and are usually plotted on coördinate paper. Measurements are made from two axes, the horizontal being called the abscissa, the vertical the ordinate. Upon these axes the respective scales are represented, that for the abscissa reading from left to right, that for the ordinate up and to the left. The independent variable is usually placed on the abscissa, the dependent occupying the ordinate axis. The stub of a series is usually the independent variable, therefore in a time series the time element is placed along the abscissa.

In the construction of curves the determination of scales is an important problem. It is self-evident that the two scales do not have to be identical, the only requisite being that equal distances on *any one axis* must represent equal changes in magnitude (absolute or relative). What distance shall represent a unit of change is largely a matter of choice and determined by the purpose of the chart. Scales should be such that the resulting curves give clearness to the data and fair proportions to the graph, the curves being neither too dispersed nor too contracted.

[1] KARSTEN, K. G., "Charts and Graphs," p. 150.

The vertical or ordinate scale should be such that the highest point of the graph, unless it represent an abnormal item, covers about two-thirds of the field of the ordinates. The field, in turn, should occupy about one-half the length of the page. Such construction gives perspicuity to the graph and allows sufficient space for an adequate title and the recording of the

POPULATION OF UNITED STATES BY DECADES, 1860–1910

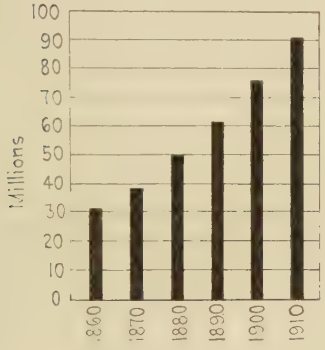


FIG. 1.

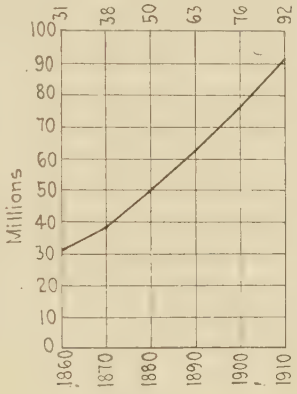


FIG. 3.

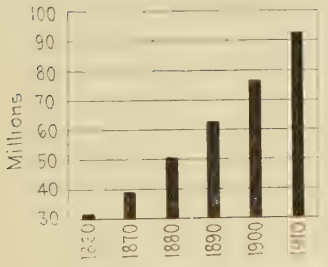


FIG. 2.

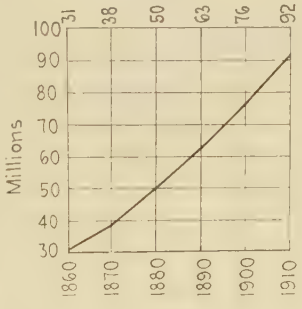


FIG. 4.

CHART 16.

data immediately above the field. Graphs, like tables, should always have titles. With the use of coordinate paper, preferably ten squares to the major ruling, the scale can easily be obtained by dividing the range of the series by the number of squares on the ordinate available for plotting. This results in a definite scale per square. A scale should however, permit the integers to fall on lines and not between lines, likewise the

five's or their multiples should be placed on major rulings (the tenth line). This arrangement gives facility in plotting.

The scale on the ordinate should indicate the zero measure of the dependent variable, otherwise erroneous conclusions may be derived from the chart. Chart 16 illustrates this need.

TABLE X

YEAR	POPULATION
1860	31,443,321
1870	38,558,371
1880	50,155,783
1890	62,947,714
1900	75,994,575
1910	91,972,226

At first glace, the reader may be disinclined to believe Figs. 1 and 2 to be representations of the data found in Table X. In the comparison of the censal years 1860 and 1870 of the two figures, the vertical bar in Fig. 2 for 1870 is approximately four times that representing 1860; this is an extraordinary increase in population. The misinterpretation is explained in Fig. 1, by means of which the population increase for the decade in question falls to about 20 per cent, which is quite a different interpretation and a more logical one. The discrepancy between the two figures is explained by the lack of the horizontal zero in Fig. 2. The curves in Figs. 3 and 4 replace the bars of the other figures.

Chart 17 is another illustration of "charts that mislead." J. Wenzel in commenting on these figures states:

> It seems hardly credible that a financial house which sells industrial securities should wilfully misrepresent the sales of one of the companies whose stock it was trying to sell. Nevertheless, if anyone saw the diagram shown in Fig. 7, he would be given an entirely erroneous impression of the difference between sales for January, 1913, as shown in the left-hand column and the sales of October, 1916, as represented by the column on the extreme right.
>
> Only after referring to the scale at the left will it be seen that, instead of starting at zero, the scale begins at 130 millions and ends with 240 millons, so that the short column at the left, which is about one-hundredth as long as the column at the extreme right, really stands for 131 millions, or more than half as much as the column at the right which stands for 238 millions.
>
> Although the volume of sales for each quarter appears in figures at the top of each column, the reader may get real facts only by a most careful examination of the scale and the figures for each quarter. The real

function of the graphic chart is to present facts so that the whole story may be told at a glance, and a hasty glance at this diagram would hardly fail to create a wrong impression.

In Fig. 8 the diagram has been redrawn so as to show the correct methods of presenting the facts . . . In this the base line begins at zero instead of 130 millions. The bar representing the January, 1913, sales is more than half as long as the bar representing the October, 1916, sales, which conforms with actual facts.[1]

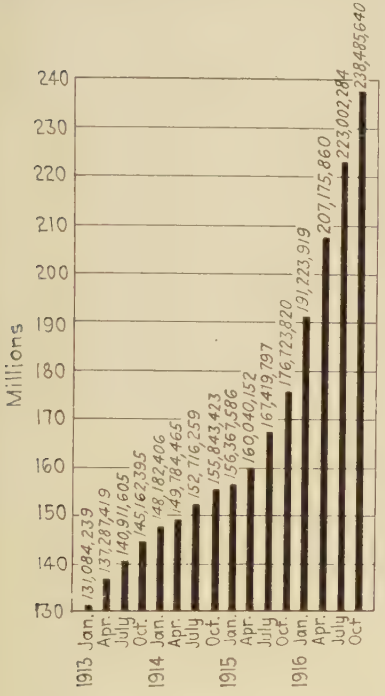


Fig. 7.

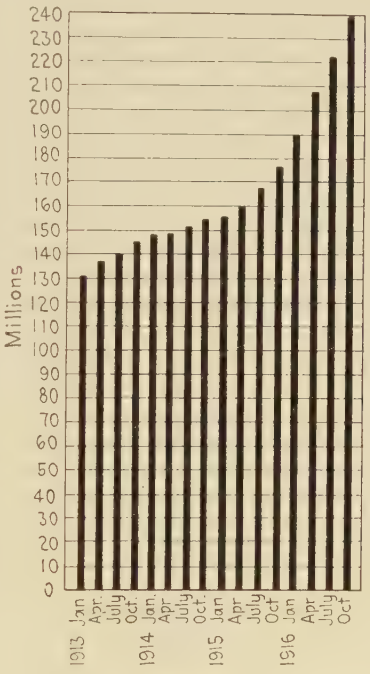


Fig. 8.

Chart 17.

36. Methods of Indicating Vertical Zero.—The vertical zero can be used when a problem is presented in which the series has a very large range. In order to include the zero, the scale for such range would be so large per square that the degree of change between some items, when plotted, would be almost imperceptible; therefore, it might be argued that such a scale should not be used. Such argument is tenable because a chart must have clearness; however, a scale may be constructed which will

[1] Wenzel, J., "Graphic Charts That Mislead," *Sci Amer. Supp.*, p. 380, June, 1917.

not only permit the inclusion of the zero but also a relatively small value per square. Chart 18 illustrates the method.

It is noticed, at once, when glancing at Chart 18, that the wave-like lines indicate that a part of the graph is missing. The zero, it will be noticed, is at the bottom left-hand corner at the intersection of the two axes. Immediately above it the ordinate breaks, continuing again at 399. Such a device prevents such

VALUE OF GOODS PRODUCED BY COÖPERATIVE SOCIETIES IN GREAT BRITAIN, 1913–1922

Source: Ministry of Labor Gazette, London, Oct., 1923

* Years 1920-1922 Do Not Include Ireland

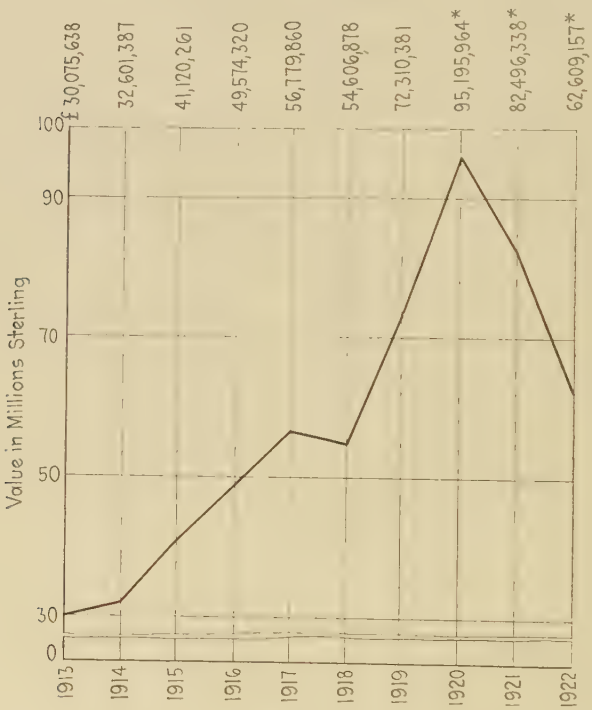


CHART 18.

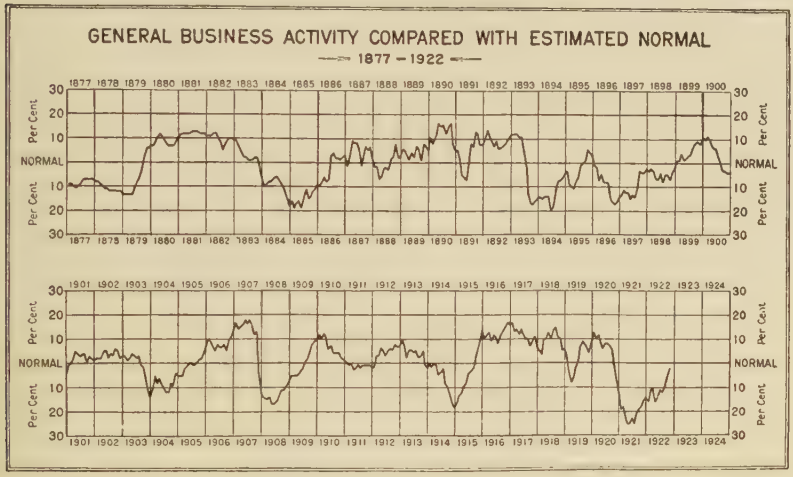
misconceptions as were indicated in Charts 16 and 17, and, at the same time, permits large calibrations. *Therefore, in every rectilinear chart it is essential for the vertical zero to be placed in the field of the chart.*

Some writers have urged the necessity of the inclusion of the horizontal zero in frequency curves. While this argument

has some justification, nevertheless, such absurd interpretations are not likely to arise by its omission.

The horizontal scale should permit ease of reading the units and give perspective to the chart. The scale, naturally, depends, to a considerable extent, upon the number of items. Items should never be crowded in the graph. The arrangement is, as already indicated, from left to right; in time series the most remote date begins at the left.

37. Placement of Figures about Axes.—When the scales have been determined they should be indicated upon the axes.



By permission of S. L. Andrew.

Chart 19.

The figures should be placed immediately to the left of the ordinate axis and immediately below the abscissa, the numbers being so placed that respective ordinates and abscissae of measurement if projected through the axes, would cut the figures in half. Such practice is illustrated in Chart 18, in which the abscissa for 50 centers on the 0 of 50, and, in the ordinate likewise for the year 1914, the 4 rests immediately below that ordinate. In time series, however, when the period is not represented as compressed into a single instant, that is, when the series is plotted by weeks or months in a year, the figures indicating the years should be written under the spaces and not under the lines (see Chart 19).

Very rarely is it good policy to write the figures of the horizontal scale at the top of the diagram. By limiting the figures to a

place below the abscissa it is possible to write the actual data above the field of the graph, by recording the actual data above the respective ordinates. This plan affords both graphic and tabular presentation and, inasmuch as graphic presentations are only approximations, the *data from which they are constructed should always accompany the graph.* The data may be indicated above the field, as in Chart 18, or may accompany the graph, as in Chart 16. Not only should the data accompany the chart in order to facilitate reference, but the source from which the data were compiled should also be indicated, usually immediately below the title.

38. Rulings for Axes.—The axes from which measurements are made should not be ruled any more heavily than the other coordinate lines, unless measurements are made from zero. *The zero line is always ruled more heavily than other lines,* except the curve itself, *which must stand in bold relief within its field.*

When the scale of the diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinate, since such a diagram does not represent the beginning or end of time.[1]

IMMIGRATION TO THE UNITED STATES, YEARS 1901–1923

Source: Monthly Labor Review, Jan., 1924

CHART 20.

In Chart 20 it will be observed that the abscissa is ruled in heavily because it is a zero line, whereas the ordinate for the year 1901 has the same thickness as other ordinates.

[1] *Preliminary Report,* Joint Committee on Standards for Graphic Presentation.

Charts which show percentage figures should have the 100 and 0 per cent lines sharply distinguished from other abscissae, since these are usually bases for comparison (see Chart 19). The Lorenz curve, because the percentage figures read from 0 to 100 on both axes, is set in a heavy frame (see Chart 44, page 130).

39. Interpretation of Curves Ruled on Rectilinear Paper.— After the axes and the scales have been determined the plotting is purely mechanical; dots are placed at the intersection of the respective ordinates and abscissae of the data. These dots, when joined, result in the curve.

The curve plotted, it may be asked: What does it mean? If plotted on rectilinear paper *it indicates trend, and not ratios.* If

IMMIGRATION INTO THE UNITED STATES, 1901–1905
(*Source: Bureau of Census*

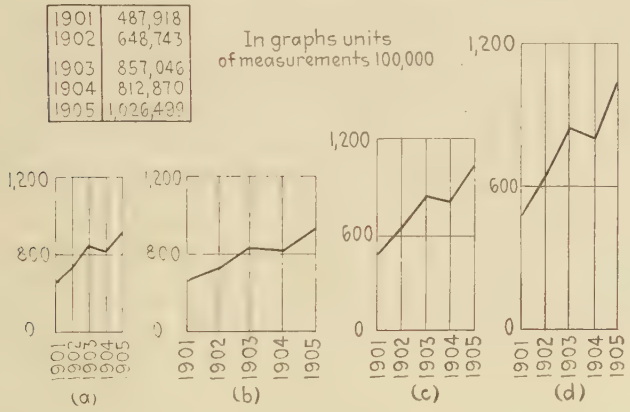


CHART 21.

the scale for the abscissa remains constant, graphs of the same data which have varying ordinate scales give very different conformations: a large scale for the latter will accentuate changes while a small one minimizes variations in the data. Conversely, with ordinate scale constant, large calibrations on the abscissa minimize, and small accentuate, changes. Chart 21 indicates how different calibrations for the same data give varying conformations to the curves.

Charts 21*a* and *b* have the same scale for the ordinate axis but are different for the abscissa. Charts 21*c* and *d* vary only in the ordinate scale. At first glance, one would hardly recognize all the curves to be representations of the same data. *Curves therefore, must be interpreted in accordance with their scales.*

The reader may wonder why all the illustrations shown are not drawn on coordinate paper of the ten-square variety. For clearness, all diagrams should be drawn on paper having only sufficient lines to indicate the scale and to guide the eye. All the illustrations, except citations, were drawn by the use of an engineer's rule, which makes possible better scale determinations than the ruled coordinate paper. For ordinary purposes, however, ten-square paper suffices.

40. Superposition of Curves.—It is often desirable to superimpose one curve upon another in order to indicate relationships,

PERCENTAGE GRAPH OF IMMIGRATION AND BUSINESS FAILURES IN UNITED STATES, 1890–1915

Based upon Table XI

Source: Dun's and Bureau of Census

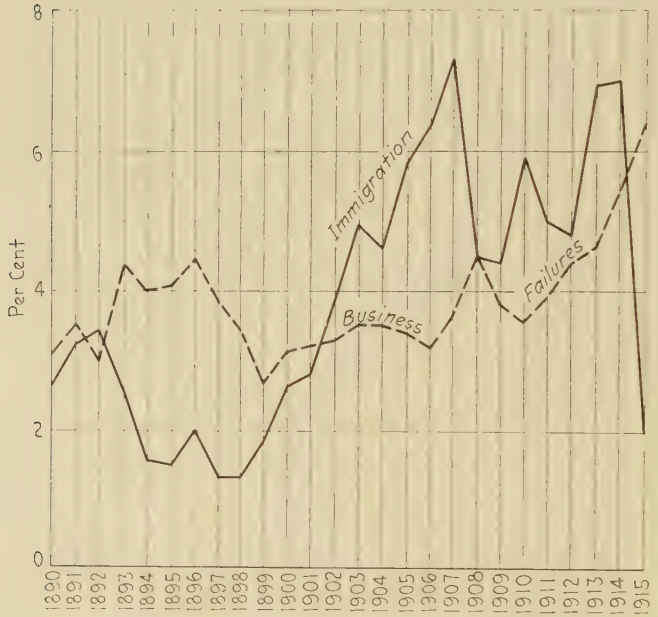


CHART 22.

but unless the scales of the respective curves have been so adjusted, it is improbable that the curves will be in proximity to one another. If it is desired to compare business failures with immigration for the years 1890 to 1915 it is evident that, if the two curves were plotted to the same scale, immigration, which has a range of over 1,000,000, would be far above the curve for failures, which has only a range of 15,000 (see Table XI).

TABLE XI.—IMMIGRATION AND BUSINESS FAILURES IN THE UNITED STATES, 1890–1915

Sources: Dun's and Bureau of Census

Years	Immigration		Business failures	
	Number	Per cent of total	Number	Per cent of total
A	B	C	D	E
1890	455,302	2.6	10,907	3.1
1891	569,319	3.2	12,273	3.5
1892	579,663	3.4	10,344	3
1893	439,730	2.6	15,242	4.4
1894	285,631	1.6	13,885	4
1895	258,536	1.5	13,897	4.1
1896	343,267	2	15,088	4.4
1897	230,832	1.3	13,351	3.9
1898	229,299	1.3	12,186	3.5
1899	311,715	1.8	9,337	2.7
1900	448,572	2.6	10,774	3.1
1901	487,918	2.8	11,002	3.2
1902	648,743	3.7	11,615	3.3
1903	857,046	4.9	12,069	3.5
1904	812,870	4.6	12,199	3.5
1905	1,026,499	5.9	11,520	3.4
1906	1,100,735	6.4	10,682	3.2
1907	1,285,349	7.4	11,725	3.6
1908	782,870	4.5	15,690	4.5
1909	751,786	4.4	12,924	3.8
1910	1,041,570	5.9	12,652	3.6
1911	878,587	5	13,441	3.9
1912	838,172	4.8	15,452	4.4
1913	1,197,892	6.9	16,037	4.6
1914	1,218,480	7	18,280	5.4
1915	326,700	1.9	22,156	6.4
Total	17,407,083	100	344,728	100

There are two ways, when using rectilinear paper, to bring these curves into juxtaposition: first, by expressing each yearly item as a percentage of the total of all items for the period; second, by the readjustment of scales.

The first method is illustrated in Chart 22, in which the percentage figures found in columns C and E of Table XI are used. This method is a laborious one and, while it reduces diverse series to the same scale, thus avoiding the difficulties already indicated in having different calibrations, it is not without its disadvantages. The chart is a plot of *relative numbers*, the absolute values being covered up in the graph.

Chart 23 illustrates the second method of bringing curves into juxtaposition. Comparison is obtained by adjusting the

Immigration into and Business Failures in the United States, 1890–1915

Sources: Dun's and Bureau of Census

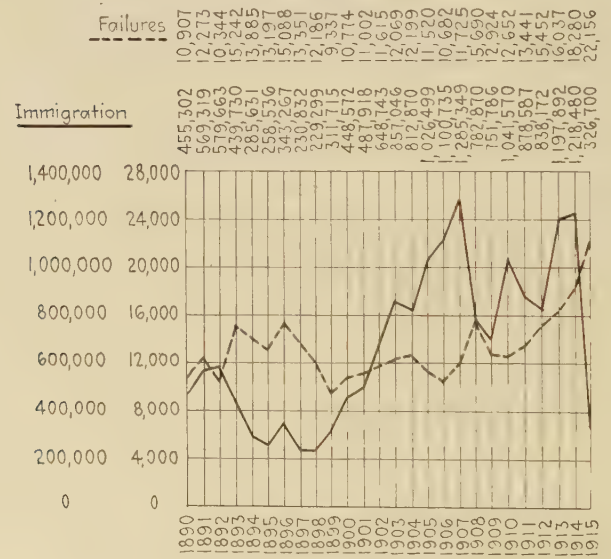
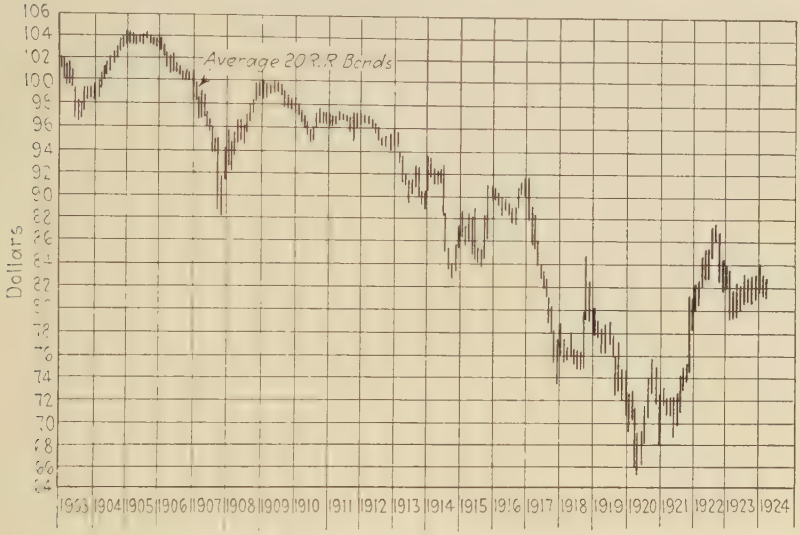


Chart 23.

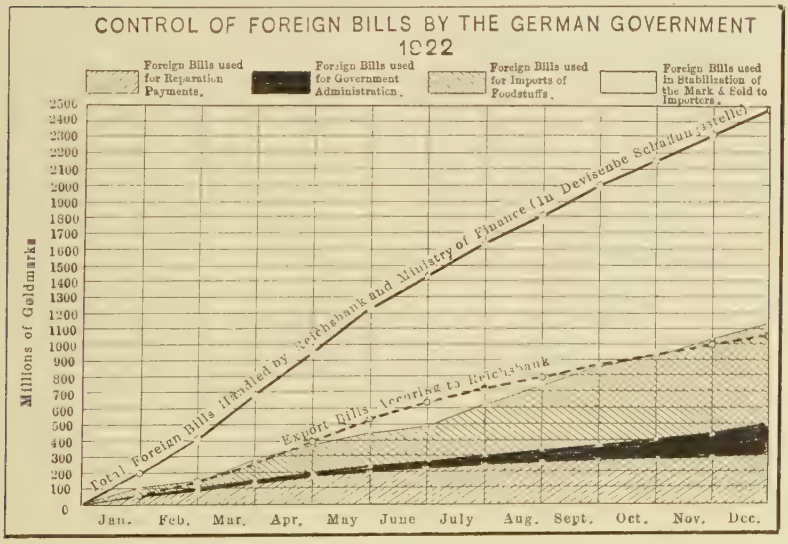
two scales, which should always be placed at the left, in such a manner as to bring the two curves closely together. In the illustration the scales are in the relation of 50 to 1. In order to differentiate one curve from another, one is plotted with a continuous line, while the other is broken. The legend indicates the series. By using dots, dashes, and their combinations many curves may be included in the same graph; however, too much information should not be crowded into any one chart, as this

detracts from its clearness. In comparing charts by this method it is well to remember that curves merely indicate trends, and because of differences in scale the slopes may vary, even though the percentage change in data is identical.



By permission of Brookmire Service.

CHART 24.



By permission of Douglas Miller.

CHART 25.

41. Zone Curve.—Another type of curve frequently used in business is the zone curve. It is used to show the zone of fluctuation in the prices of goods sold on the produce and cotton exchanges and in large markets, and the fluctuations in the price of stocks and bonds. Chart 24 indicates the monthly high and low quotations of twenty-two railroad bonds.

42. Band Diagram.—At the beginning of this chapter it was suggested that for component series the pie diagram or 100 per cent bar graph could be used. Such a diagram, however, is restricted to one period of time—it cannot be used to show the component parts over a series of years. The band diagram (Chart 25) is used for such presentations. This is a very useful type of diagram, and by the process of cross-hatching the many constituent parts of the whole may be designated in such a way that the graph may contain a wealth of information.

43. Logarithmic Charts.—*a. Description.*—Comparison, however, no matter what type of curve, is a hazardous thing because,

> a curve of uniform slope on a chart of rectilinear lines indicates only that there has been a uniform increase or decrease in actual numbers, not in uniform rate of change on a percentage basis.[1]

Because of the difficulties in reading the ratio of change in the curves plotted on rectilinear paper it is better to use logarithmic paper. Logarithmic paper, because of its many advantages and the simplicity of its use, is now being extensively used.

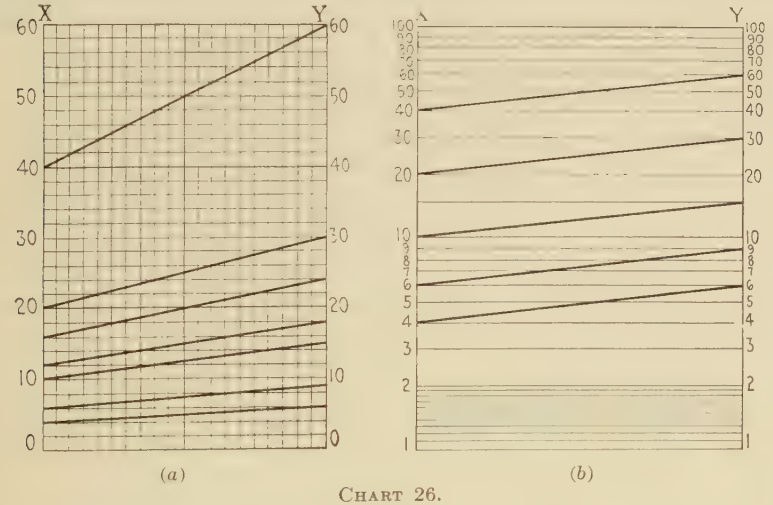
The horizontal rulings on logarithmic paper are scaled in such a way that the spaces are not in proportion to the numbers but, rather, to the logarithms of those numbers—in other words, the spacing is identical with that on a slide rule. The vertical rulings are, usually, in arithmetic spacing. The introduction of the term "logarithm" may sound formidable to the student, however, a knowledge of logarithms is not required in order to understand the use of the paper. The paper may be purchased already ruled in one, two, or three cycles. All that is required of the student is the ability to use and interpret the paper.

b. Why Used.—Before illustrating how the paper differs from the ordinary coordinate paper, it might be well to give some of the reasons for advocating its use.

[1] Bowley, A. L., "Elements of Statistics," p. 73.

With ordinary arithmetical scale, fluctuations in large factors are very noticeable, while relatively greater fluctuations in small numbers are barely apparent. Logarithmic scale permits the graphic representation of change in every quantity without respect to the magnitude of the quantity itself. At the same time, the logarithmic scale shows the actual value by reference to the numbers in the vertical scale. By indicating both absolute and relative values and changes, the logarithmic scale combines the advantages of both natural and percentage scales without the disadvantages of either.[1]

The quotation should indicate the need for mastering the use and application of logarithmic paper.



(*a*) (*b*)

CHART 26.

Charts 26*a* and *b*, which are based on Table XII, illustrate the differences in the ruling of rectilinear and logarithmic paper.

TABLE XII.—TABLE OF PAIRED ITEMS

Ordinates	
X	*Y*
4	6
6	9
10	15
20	30
40	60

[1] CUNNINGHAM, W. J., "The Logarithmic Scale in Graphic Charts," *Ry. Age Gaz.*, p. 1518, June 25, 1909.

Chart 26*a* shows the ordinary coordinate paper, in which each space on the vertical represents two units of measurement. The distance from 10 to 20 is identical with that from 20 to 30, 50 to 60, etc. The ratio of change, however, is not identical, as is shown in Table XIII. The rectilinear graph therefore hides very effectively the visualization of ratio relationships. The lines representing the data, it should be noticed, all diverge from one another.

TABLE XIII

5 spaces measure	100	per cent change from 10 to 20
	50	per cent change from 20 to 30
	33	per cent change from 30 to 40
	25	per cent change from 40 to 50
	20	per cent change from 50 to 60
	16+	per cent change from 60 to 70
	14+	per cent change from 70 to 80
	12+	per cent change from 80 to 90
	11+	per cent change from 90 to 100

c. The Derivation of Logarithmic Rulings.—Chart 26*b* illustrates the same material as *a*. The vertical scale, however, is logarithmic. It is derived by spacing according to the *logarithms* of the numbers 1 to 10. There is no zero line in logarithmic paper, because no matter how small the differences between

TABLE XIV

Number A	Logarithms B
1	0.0000
2	0.3010
3	0.4771
4	0.6021
5	0.6990
6	0.7782
7	0.8451
8	0.9031
9	0.9542

figures may be the percentage can never be 100; that is, by taking a constant percentage from a number, the number decreases but never reaches zero. In order to illustrate the principle involved in the use and construction of logarithmic paper the logarithms of the numbers are given in Table XIV.

The spacings on the logarithmic paper are made according to column B of Table XIV. They may readily be made by the use of an engineer's rule. The space from 1 to 2 would equal 0.3010 gradations on the rule, from 1 to 3, 0.4771 gradations, etc. By continuing up to 9 the scale would be completed. It can be repeated above 9 according to the number of cycles required. The intermediate measurements, such as halves, would be derived in a similar manner.

d. Differences from Rectilinear Paper.—The principle of such paper may perhaps best be observed by the following table. which is based upon the logarithms in column B, Table XIV,

Table XV

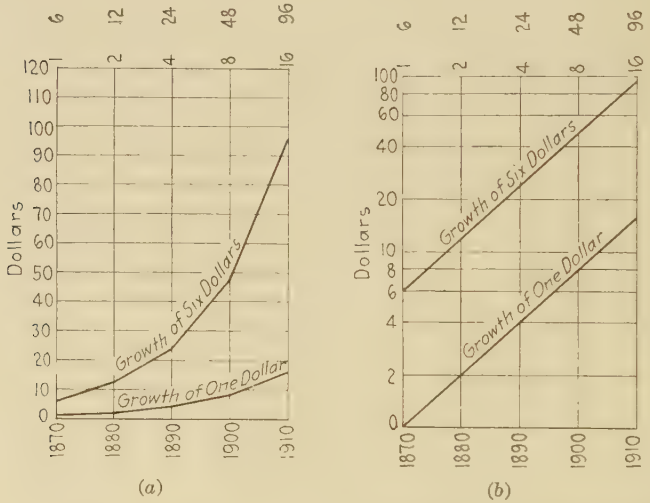
Numbers A	Logarithms of numbers B	C
2–1	0.3010 − 0.0	= 0.3010
4–2	0.6021 − 0.3010	= 0.3011
8–4	0.9031 − 0.6021	= 0.3010
3–1	0.4771 − 0	= 0.4771
9–3	0.9542 − 0.4771	= 0.4771
3–2	0.4771 − 0.3010	= 0.1761
6–4	0.7782 − 0.6021	= 0.1761
9–6	0.9542 − 0.7782	= 0.1760

From this table it is evident that equal percentage of change is represented by equal distances, irrespective of the change in absolute numbers. Column B lists the scale for the various numbers, while column C indicates the linear distance between such numbers. The latter column shows that the distance from 2 to 3 is identical with those from 4 to 6 and from 6 to 9, since all are represented by 0.1761 gradations; likewise, for the other

groupings. From the above the first rule may be deduced: *Equal percentage change is always represented by the same vertical distance on the scale.* Therefore, in Chart 26*b* the lines representing the data are parallel in contrast to the diverging lines of *a*. This indicates that

> the ordinary chart is adapted to difference comparisons rather than to ratio comparisons, whereas the statistician is usually concerned with ratio comparisons far more than difference comparisons.[1]

The validity of the contention that ratio charts should be used is well illustrated by Chart 27. In this chart there appears to be very little relationship between the curves in *a* and *b*, whereas they are plotted from the same data. *a*, however, is drawn to arithmetic scale, while *b* employs the logarithmic. It should be observed that the logarithmic scale has two cycles; it is



(*a*) (*b*)

From Wenzel, J., *Graphic Charts, Scientific American, p.* 236, *April* 14, 1917.

Chart 27.

capable, within its field, of representing data with a hundred fold increase. In *b* the two curves run parallel to each other, thus showing equal percentage of change, or, putting this in the form of a rule: *Curves having the same percentage change, irrespective of absolute numbers, run parallel to one another.*

[1] Fisher, Irving, "The Ratio Chart," *Amer. Stat. Pub.*, p. 578, June, 1917.

On the arithmetic scale fluctuations in large numbers are exaggerated relatively to those in smaller number. Charts 28 and 29 are representations of the same material. The charts, however, are ruled in arithmetic and logarithmic scale, respectively.

In comparing the curves in Chart 28 it should be evident that

> a serious defect in the difference method is that in a curve of rapid growth the difference chart is useful only in the middle portion. At the extreme left . . . it becomes almost indistinguishable from a horizontal line and at the extreme right it becomes almost indistinguishable from a vertical line. At either extreme no eye can estimate the percentage rate of growth, although the rate may not be different from the rate in the middle.[1]

Chart 28 conveys the impression that the growth of pig iron was almost imperceptible in the early decades, but culminated in rapid expansion in the more recent decades. Such an impression is totally misleading, as is indicated by Chart 29, which shows that the percentage growth of pig-iron production in the early censal years was relatively greater than that of the later

PIG IRON PRODUCTION AND IMMIGRATION THE UNITED STATES AT CENSAL YEARS, 1830–1920

Data and Sources on Chart 29

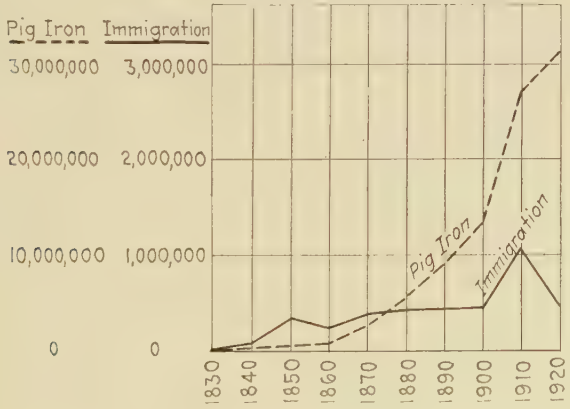


CHART 28.

period. This fact is indicated by the steeper inclination of the curve in the earlier year in contrast to the later years.

By comparing the two curves on the rectilinear Chart 28 for the decades 1860 to 1870 and 1900 to 1916 it appears that pig-

[1] *Ibid.*, pp. 585–586.

iron production increased faster than immigration in both periods. This, too, is a false interpretation. Logarithmic Chart 29 gives a sharper inclination to the immigration curve than that for pig-iron production for the decades mentioned, thus reversing the interpretation of the arithmetic chart. From this fact another rule may be inferred: *The vertical rise of a curve on logarithmic paper indicates the percentage change, the steeper the inclination the greater the percentage.*

Pig Iron Production and Immigration the United States at Censal Years, 1830–1920

Source: Bureau of Census

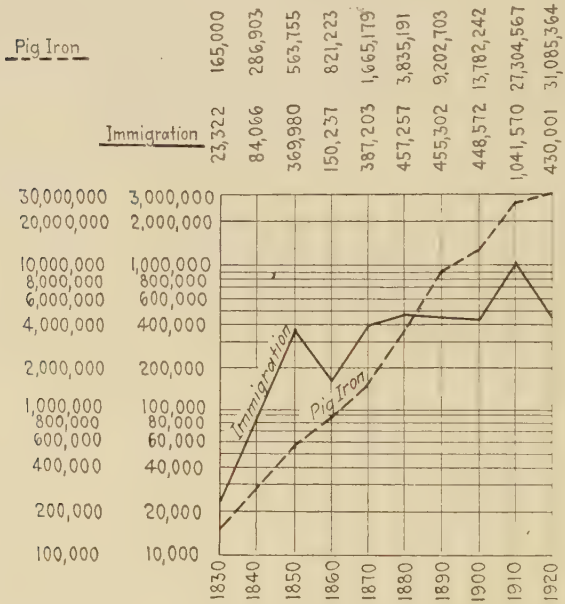
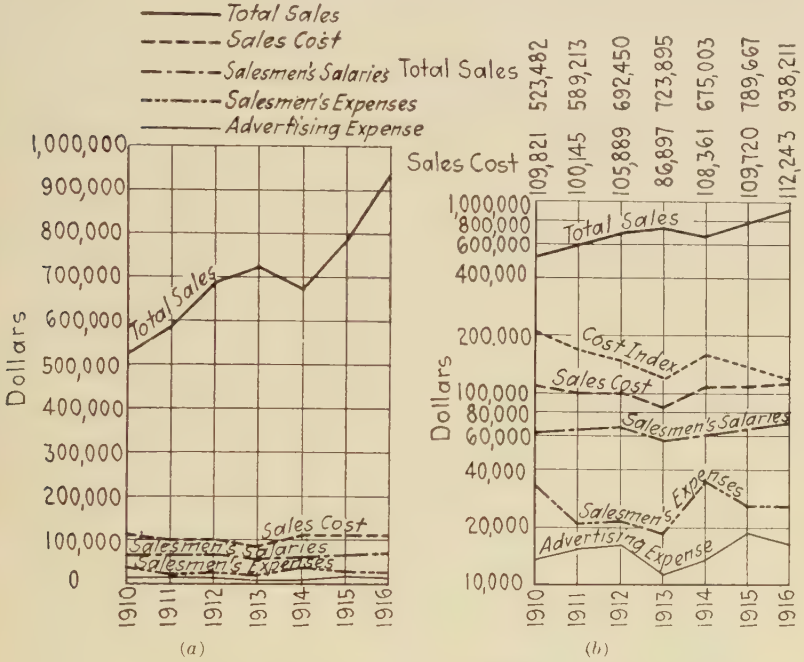


Chart 29.

In using a logarithmic chart there is little need for the reduction of absolute numbers to relative numbers, as illustrated in Chart 22, in order to bring curves into juxtaposition. By scale adjustment the ratio chart may retain the absolute numbers, permit curves to be superimposed, and, at the same time, indicate ratio of change. Chart 29 indicates how curves may be brought into proximity to one another by scale adjustment. Because the ratio chart has no zero line it is well to start the scale

with some power of 10. This chart has two scales, but each starts and finishes with a multiple of 10. Another observation made in comparing the two (Charts 28 and 29) is that the ordinate scales on both are in the same proportion, 10 to 1. Nevertheless, the graphs show vastly different dispersions. The



Year	Total Sales	Sales Cost	Salesm'n Salary	Salesm'n Expense	Adv. Expense	Cost Index
1910	$523,482	$109,821	$63,259	$33,002	$13,560	21
1911	589,213	100,145	65,212	20,830	15,103	17
1912	692,450	103,889	66,490	21,490	16,212	15
1913	723,895	86,897	57,211	18,556	11,130	12
1914	675,003	108,361	60,576	34,575	13,445	16
1915	789,667	109,720	65,110	26,021	18,589	14
1916	938,211	112,243	70,364	25,647	16,232	12

WENZEL, J., *Graphic Charts, Scientific American, p. 237, April 14, 1917.*

CHART 30.

ratio chart gives the smaller series a better proportioned field than the rectilinear, as is indicated by the area covered by the immigration curve in the respective graphs.

With logarithmic paper, tables having great range may be plotted with ease because the paper may be made in as many

cycles as necessary. On the three cycles in Chart 29 the range is from 100,000 to 30,000,000 and yet the curve is given in good perspective.

As a further illustration of the usefulness of the ratio chart, Chart 30 is given, in which *a* shows the meager information given on rectilinear paper and the wealth contained when logarithmic paper is used.

e. Interpretation of Slope on Logarithmic Paper.—By this time the student should be willing to agree with Cunningham that logarithmic paper "combines the advantages of both natural and percentage scales without the disadvantages of either." Irving Fisher[1] summarizes the ways of utilizing the direction in ratio charts as follows:

1. If we see a curve ascending, and nearly straight, we know that the statistical magnitude it represents is increasing at a nearly uniform rate.
2. If the curve is descending, and nearly straight, the statistical magnitude is decreasing at a nearly uniform rate.
3. If the curve bends upward, the rate of growth is increasing.
4. If downward, decreasing.
5. If the direction of the curve in one portion is the same as in some other portion it indicates the same percentage rate of change in both.
6. If the curve is steeper in one portion than in another portion, it indicates a more rapid rate of change in the former than in the latter.
7. If two curves on the same ratio chart run parallel they represent equal percentage rates of change.
8. If one is steeper than another the first is changing at a faster percentage rate than the second.
9. The imaginary straight line most nearly representing, to the eye, the general trend of the curve is its growth axis, and represents the average rate of increase (or decrease); and the deviations of the curve from this growth axis are plainly evident without recharting.
10. The slope of the imaginary line, between any two points, on a curve indicates the average rate of change between the two.

From these rules it may be inferred that direction in logarithmic charts may be used in forecasting probable future growth.

In proportion to the space devoted to other graphic methods it may appear that too much space has been given to logarithmic charts. They are of sufficient importance, however, to be given the major portion of the discussion on curves. It should be

[1] *Ibid.*, pp. 597–598.

clearly understood that this type of chart is easy to understand and its simplicity and accuracy should command more attention from chart makers.

44. Summary.—In this chapter it may be noticed that no attention has been given to the graphic presentation of series measuring variable phenomena; frequency series. Such an omission has been made because frequency curves have special problems which can only be understood after a general discussion of the problems of frequency series. This discussion has been reserved for Part II.

No summary is necessary in concluding this chapter. The student should be impressed by the facts that graphs aid in analysis and are not really analysis; they may afford the basis for pursuing further analysis but, at best are only approximations. In so far as they are essential for illustration great care should be taken in their construction, in order that truthful pictorial presentation of data may be given.

References

Bailey, W. B.: "Modern Social Conditions," pp. 54–56.

Bowley, A. L.: "An Elementary Manual of Statistics," pp. 35–50.

Brinton, W. C.: "Graphic Methods for Presenting Facts."

Cunningham, W. J.: "The Logarithmic Scale in Graphic Charts," *Ry. Age. Gaz.*, June 25, 1909.

Fisher, Irving: "The Ratio Chart," *Quart. Pub. Amer. Stat. Assoc.*, June, 1917.

Karsten, K. G.: "Charts and Graphs."

Kelley, Truman L.: "Statistical Method," pp. 37–43.

King, W. I.: "Elements of Statistical Method," chap. X.

Knoeppel, C. E.: "Graphic Production Control."

Secrist, H.: "An Introduction to Statistical Methods," chap. VI.

——: "Readings and Problems in Statistical Methods," chap. VI.

——: "Statistics in Business," chap. VI.

Warne, F. J.: "Chartography."

Wenzel, J.: "Charts That Mislead," *Sci. Amer.*, June, 1917.

——: "Graphic Charts," *Sci. Amer.*, April, 1917.

West, Carl: "Introduction to Mathematical Statistics," chaps. I–II.

Whipple, G. G.: "Vital Statistics," chap. III.

Problems

1. Given the following table:

POPULATION AND AREAS OF THE CONTINENTS AND COUNTRIES OF THE WORLD, 1920

Source: Statesman's Year Book, 1920

Region or country	Area in square miles	Population	Per cent distribution	
			Area	Population
	A	B	C	D
Aggregate	55,885,000	1,720,000,000	100.0	100.0
Continental divisions:				
Asia	17,052,000	890,000,000	30.5	51.7
Europe	3,821,000	475,000,000	6.8	27.6
North America	8,040,000	145,000,000	14.4	8.4
South America	7,018,000	61,000,000	12.6	3.5
Africa	11,605,000	140,000,000	20.8	8.1
Australia and Oceania	3,457,000	9,000,000	6.2	0.5
Polar Regions	4,892,000	.....	8.8	.....

a. Construct a pie diagram for the percentage distribution of area, column C, and one for population, column D.

b. Construct 100 per cent diagrams for columns C and D.

c. Construct a multiple horizontal bar diagram to show the relationship between the population and area for the respective divisions.

2. Given the following table:

PERSONS ENGAGED IN MANUFACTURING INDUSTRIES 1919, 1914, AND 1909

Source: Fourteenth Census Reports on Population

Class	Census year	Total	Male	Female	Per cent of total	
					Male	Female
		A	B	C	D	E
All classes	1919	10,812,736	8,583,720	2,229,016	79.4	20.6
	1914	8,263,063	6,613,376	1,649,687	80.0	20.0
	1909	7,678,578	6,162,263	1,516,315	80.3	19.7

Draw.

a. Simple horizontal bar diagram of data (column A).

b. Simple vertical bar graph of same data.

c. Horizontal composite or compound bar diagram of data in columns B and C.

d. Vertical bar of same data as in (*c*).

e. One-hundred per cent horizontal bar for columns D and E.

f. One-hundred per cent vertical bar graphs for data in columns D and E.

g. Curves for data in columns B and C. Plot curves in the same field.

h. Band diagram for data.

i. Curves for data in columns B and C, in same field, but with logarithmic ruling on the ordinate axis.

3. What interpretation does the logarithmic chart (*i*) give which the rectilinear ruling (*g*) hides or distorts? Why?

4. For a period of twenty days record from the newspaper the daily high and low quotations of a specific future in cotton sold in the New Orleans market. From these figures construct a zone chart.

5. Given the following table:

POPULATION OF THE UNITED STATES, BY RACE OR COLOR, 1890, 1900, 1910, AND 1920

Source: Bureau of Census

Census year	Race or color		
	Total population A	White B	Indian C
1920	105,710,620	94,820,915	244,437
1910	91,972,266	81,731,957	265,683
1900	75,994,575	66,809,196	237,196
1890	62,947,714	55,101,258	248,253

Plot the two curves for the data in columns B and C in the same field as follows:

a. On rectilinear paper in such a way as to bring the two curves into juxtaposition.

b. On logarithmic paper, with the use of only one scale.

6. From the illustrations above summarize the advantages of logarithmic paper.

PART II

STATISTICAL METHODS FOR FREQUENCY SERIES

CHAPTER VII

CLASSIFICATION AND TABULATION

45. Introduction.—Statistical studies may involve the measurement of variables possessing certain characteristics, which may differ in the degree with which the attributes are possessed: (1) at any one time, (2) with change over a period of time. A variable is that which changes or varies in measurement. The first type of study is that of a frequency series and is illustrated by wage studies, the distribution of wages in a factory, output and its relation to wage scales, wage scale and cost per unit of output—all these studies imply measurement of variables, *at a given time*. The second type of series typified above is the time series, as could be illustrated by a comparative study of wages over a period of time, cost per unit of production over a period of time, and the lessening of direct labor cost by improved methods of production—all these involve time and are called time series.

These two types of series are those most commonly treated in economic studies, and each has specific problems. Therefore, in this text they will be taken up separately. In this part of the book only those problems pertaining to frequency series. will be discussed. Part III will be devoted to time series.

46. Definition of Frequency Table.—A frequency table is the recording into homogeneous groups, arranged in ascending order, the occurrence of various measurements of certain phenomena which are measured at one definite period of time. Such definition does not prevent the variable being time itself, as frequency series are made which measure the hours of labor, duration of strikes, duration of disabilities, etc. In these series, however, it is apparent that the time data which are found on the stub of the table are for a definite period, usually one year, and do not involve comparison of change in the *same phenomena* during a period of time. The groups into which the data are recorded are called "class intervals," and the number of occurrences in any one class constitutes its frequency. The total of all fre-

quencies constitutes the distribution of the entire series and is termed "total frequency," or "population."

In order to bring out the problems encountered in the classification of frequency series, and, at the same time, to make more apparent what such series are, let it be assumed that the data of a wage study have been classified in ascending order, as indicated in Table XVI.

TABLE XVI.—WEEKLY RATES OF WOMEN MESSENGERS AND BUNDLERS IN RETAIL STORES IN MASSACHUSETTS

Hypothetical distribution based on *Seventh Annual Report*, Minimum Wage Commission of Massachusetts

10 workers at $5.00	20 workers at $8.00	12 workers at $10.00
5 workers at 6.00	40 workers at 8.25	14 workers at 10.50
5 workers at 6.50	5 workers at 8.50	3 workers at 11.50
4 workers at 6.75	16 workers at 8.75	1 worker at 12.25
15 workers at 7.00	10 workers at 9.00	3 workers at 12.50
15 workers at 7.50	5 workers at 9.50	1 worker at 12.75
5 workers at 7.75	5 workers at 9.75	1 worker at 13.50
		1 worker at 15.50

47. An Array.—These data may be represented graphically, and thus aid in their analysis, by means of an array. An array is the graphic presentation of data by means of ordinates, arranged in ascending order, the length of the ordinate representing magnitude or size, each datum being represented by a separate ordinate. The application of this method to the data of Table XVI is shown in Chart 31.

In glancing at the array the reader's attention is at once arrested by its step-like appearance. The data cluster in groups, showing that there is a certain homogeneity within certain groups of items. The array, however, like the table from which it was constructed, gives far too much detail, and this form of graphic presentation conveys little more to the reader than the data themselves. The question then arises: How may some of the detail be sacrificed? The array should suggest the answer: Why not increase the rise in the steps, that is, why not break the data up into groups having wider limits? All the data are wage data (weekly wages). The one attribute wages is common to all. But wages may be classified into groups having an additional homogeneous attribute—the magnitude of wages. Weekly

wages are rarely paid in units less than 25 cents. The grouping, therefore, may be some multiple of this sum. In this problem let the grouping be an interval of $1. By tabulating the classified data of Table XVI into $1 groups, making no regular intervals below $7, Table XVII is constructed.

A frequency table is the condensation of the classified data having one common attribute (wages) into groups having not

AN ARRAY OF DATA OF TABLE XVI

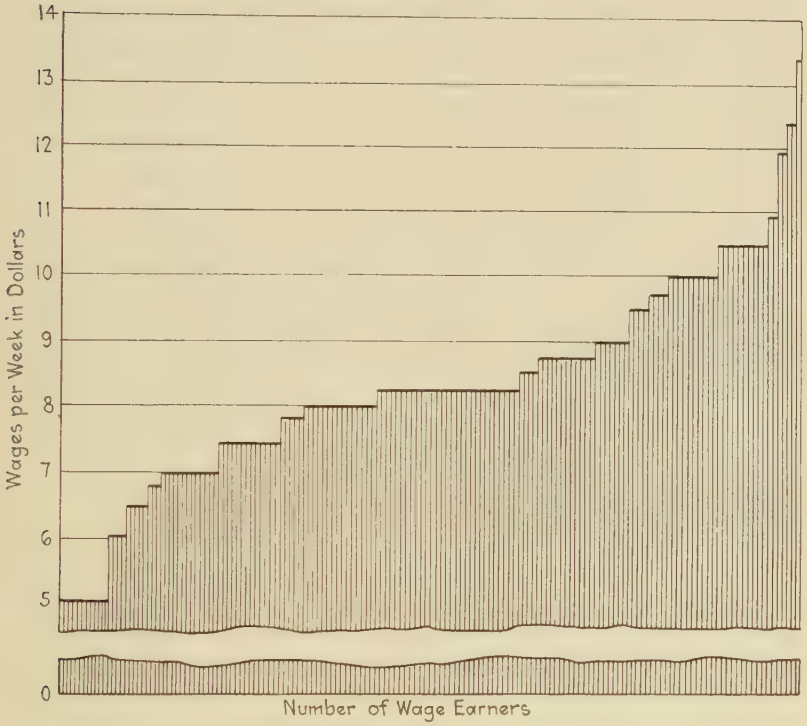


CHART 31.

only this attribute but certain homogeneous qualities with regard to magnitude (rate of wages). In Table XVII, thirty-five items are recorded in the group "7 and under 8", which shows that thirty-five workers received wages within these *limits*. Table XVII presents the data in a more effective way than either the simple classification of Table XVI or the array of Chart 31. It should be evident, however, that this classification into homogeneous groups sacrifices much detail. The student cannot tell how many of the thirty-five women within the "$7 and under

TABLE XVII.—WEEKLY WAGES OF WOMEN MESSENGERS AND BUNDLERS IN RETAIL STORES IN MASSACHUSETTS

Source: Seventh Annual Report, Minimum Wage Commission of Massachusetts (page 40)

Wages per week	Employees
Under $ 7	24
$ 7 under 8	35
8 under 9	81
9 under 10	20
10 under 11	26
11 under 12	3
12 under 13	4
13 under 14	1
14 under 15	0
15 under 16	1
Over 17	1
Total	196

$8" class received the wage of $7.25; as a matter of fact, by consulting Table XVI, it is seen that no such wage was paid. Important problems, therefore, are involved in the grouping of data for frequency series.

The problems of grouping are important inasmuch as the resulting table sacrifices detail; the classification should be such that it will not sacrifice, or include, too much detail. Naturally, the amount of detail varies with the width of the class intervals, the wider the interval the less the detail, and *vice versa*. The first problem, then, confronted in the classification of frequency series is the determination of the width of the class interval.

As suggested above, the interval classification prevents the visual derivation of the frequency for any value (interpolation) within the class. Some point must, therefore, be chosen which is considered representative of all data within the group. This point must of necessity be the midpoint of the class, and should be the point of normal distribution (the data concentrating regularly on either side of this point). The midpoint, however, varies according to the limits of the class. A class, for instance,

indicated as $8 and under $9 would have a midpoint of $8.50, whereas, if it were indicated $8 to $9, the data being measured to the nearest 50 cents, the limits would be $7.75 to $8.75 with a midpoint of $8.25. Therefore, not only is there a problem in the determination of the interval, but also one connected with the class limits. These two problems must of necessity be solved before data may be classified into a frequency table.

Before analyzing the problems of class intervals and limits it is necessary to explain two terms, continuous and discrete, which arise continually in the discussion of frequency series. The terms are connected with the unit of measurement.

48. Continuous Series.—A continuous series is one in which the phenomena, if sufficient in quantity and measured with utmost accuracy, would mass at a norm (point of highest frequency); above and below this norm would be found data differing in size by the smallest gradations, so that if the data were plotted in the form of an array the top of the ordinates would present a curve-like appearance. Table XVIII illustrates continuous data. Chart 32, constructed from Table XVIII, shows the curve-like appearance of the data. In this array there is no step-like appearance—the ordinates increase gradually. This array explains the significance of the term "continuous"—there are all gradations in measurement, one seeming to spring from another. When such series are thrown into tabular form with wide class intervals, the class intervals and midpoints become purely arbitrary designations. Table XIX illustrates the classification and tabulation of the material in Table XVIII. In this table the data appear to concentrate in the group "50 and under 60" which has a frequency of 13. Arbitrarily, such grouping infers that fifty-five years could designate all frequencies within the class. This measurement, however, is an approximation, it is possible for a frequency to occur at any year, month, day, etc. within the group. The series then is continuous, when an arbitrary measurement of one year is not adopted.

Most series involving the measurement of natural phenomena are continuous, such as linear measurements; no matter how accurately the phenomena have been measured it is still possible to get more accurate calibrations, which in turn would break up groups which were supposedly homogeneous in size.

TABLE XVIII.—AGE OF WORKERS IN DEPARTMENT Y OF FACTORY X

Number of workers	Age	Number of workers	Age	Number of workers	Age
1	33	1	49	1	61
2	35	1	50	1	62
1	37	1	51	1	63
1	40	1	52	1	64
1	41	1	53	1	65
1	42	2	54	1	66
1	43	2	55	1	67
1	44	2	56	1	68
1	45	1	57	1	69
1	46	1	58	1	75
1	47	1	59	2	85
1	48	1	60		

CHART BASED UPON TABLE XIX

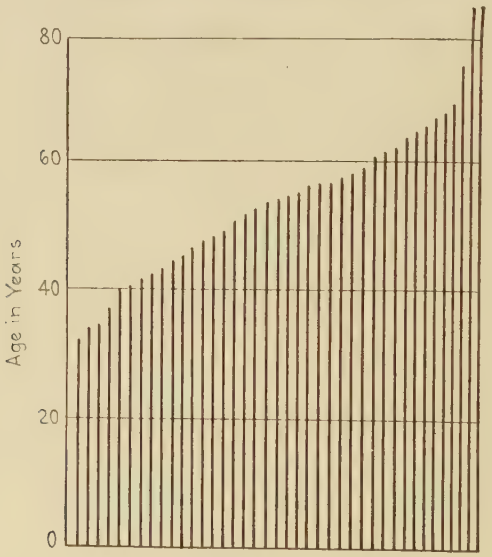


CHART 32.

Continuous series, therefore, are those in which the measurements are only approximations and are capable of further refinement. Measurements may therefore be interpolated for within classes, and the graphic presentation may be represented by a smoothed curve.

TABLE XIX.—AGE OF WORKERS IN DEPARTMENT Y OF FACTORY X

Age in years	Number of workers
30 and under 40	4
40 and under 50	10
50 and under 60	13
60 and under 70	10
70 and under 80	1
80 and under 90	2
Total	40

49. Discrete Series.—A discrete series is one in which the data are of such a character that their measurement is not capable of indefinite refinement. The change from the ordinary clover leaf to the supposedly lucky one consists of the addition of an extra leaf, and no other change in the number of leaves would be possible unless it be a multiple of 1. In the tossing of a coin, the result is definite and fixed, the only possibility being either head or tail. Weekly, monthly, and yearly wage rates are rarely paid in odd cents; arbitrary divisions of the dollar are usually the minimum increase. Such data, when plotted in the form of an array, present a step-like appearance, as will be evident by reference to Chart 31. In this array the ordinates, in contrast to those in Chart 32, present a step-like appearance. If the data are grouped into wider intervals it is illogical to assume that the class frequency distributes itself regularly over the group. In Table XVII it should not be assumed that the thirty-five items within the class \$7 to \$8 are spread over the entire interval. As already indicated *above*, no frequency is found for the \$7.25 wage; likewise no wages are expressed in odd cents, 25 cents being the smallest difference between wages. *The nature of the unit therefore determines the groupings.*

Data which usually may be considered continuous may, by definition of unit, become discrete. Age may be measured with a considerable degree of refinement, being essentially a continuous series, but for statistical purposes age recorded to the nearest year suffices. Economic data tend to be discrete, care should therefore be taken in the tabulation and analysis of frequency series. In a discrete series, if the range is not very great, the

unit of change (height of the step in the array) may be the size of the interval.

In summarizing the differences between continuous and discrete series it may be said that measurements in continuous series are only approximations and are capable of further refinement, the data differing by the smallest gradations; discrete series present data the measurement of which is definite, there being no possibility of intermediate values between the base unit.

This seeming digression has an important bearing on the problems of the class interval and limits which may be discussed.

50. The Class Interval.—The first step in the determination of the width of the interval involves the decision as to the amount of detail to be included in the table. Too large an interval distorts the distribution, while too small an interval may fail to give perspicuity to the data. As an example of too large an interval, which indicates how such a method hides important detail, the following table has been *adapted* from "Income in the United States," by the National Bureau of Economic Research.

TABLE XX.—INCOME IN THE UNITED STATES, 1918
Adapted table

Income class	Number of persons
$ 0– $ 5,000	36,726,602
5,000– 10,000	587,824
10,000– 25,000	192,062
25,000– 50,000	41,119
50,000– 100,000	14,011
100,000– 200,000	4,945
200,000– 500,000	1,976
500,000– 1,000,000	369
1,000,000 and over	152
Total	37,569,060

In Table XX, obviously, the use of an interval of $5,000 for the first few classes is incorrect, because within the limits of the first class interval is contained 97 per cent of the items in the total distribution. From the student's knowlege of income

figures it will be apparent that of the 97 per cent of income receivers the greater percentage (91) will be found below the midpoint of the 0 to 5,000 class. The prime object of many income studies is usually to measure the distribution of the smaller incomes which the above table does not show. By breaking the data into smaller intervals much more detail may be shown, as indicated in Table XXI.

TABLE XXI.—DISTRIBUTION OF INCOME IN THE UNITED STATES, 1918
Source: Income in the United States (page 136)

Income class	Number of persons
Under zero	200,000
$ 0– $ 500	1,827,554
500– 1,000	12,530,670
1,000– 1,500	12,498,120
1,500– 2,000	5,222,067
2,000– 3,000	3,065,024
3,000– 5,000	1,383,167
5,000– 10,000	587,824
10,000– 25,000	192,062
25,000– 50,000	41,119
50,000– 100,000	14,011
100,000– 200,000	4,945
200,000– 500,000	1,976
500,000– 1,000,000	369
1,000,000 and over	152
Total	37,569,060

By permission of National Bureau of Economic Research.

The first rule with regard to intervals is that they should be so chosen as to give sufficient detail for the correct interpretation of the data. An arbitrary rule, as suggested by G. U. Yule, is that there should be between fifteen and twenty-five intervals in a table. As indicated, this rule is *arbitrary*, and the number must be governed largely by the data. Where the samples are numerous and the variations large the range should be divided into some such number of intervals. Naturally, in series in which the measurements increase in jumps (discrete series) the interval can never be less than the unit itself: in Table

XVII the unit cannot be less than 25 cents; in the throwing of dice the difference can never be less than 1. Therefore, if variations are small the unit may become the interval itself.

The class interval should mass the data in such a way as to prevent many intervals of zero frequency. A repetition of zero distributions within classes indicates that either the limits of the class have been poorly chosen, or the interval is too small, Table XXII indicates how too small an interval gives too much detail, and the unnecessary repetition of zero classes, there being twenty-one such in a total of forty-three. The wages given in this table represent the midpoint of the classes. The zero intervals are smoothed out by the larger interval of Table XVII which has

TABLE XXII.—WEEKLY RATES OF WOMEN MESSENGERS AND BUNDLERS IN RETAIL STORES

Wages	Workers	Wages	Workers
5.00	10	10.25	0
5.25	0	10.50	14
5.50	0	10.75	0
5.75	0	11.00	0
6.00	5	11.25	0
6.25	0	11.50	3
6.50	5	11.75	0
6.75	4	12.00	1
7.00	15	12.25	0
7.25	0	12.50	3
7.50	15	12.75	7
7.75	5	13.00	0
8.00	20	13.25	0
8.25	40	13.50	1
8.50	5	13.75	0
8.75	16	14.00	0
9.00	10	14.25	0
9.25	0	14.50	0
9.50	5	14.75	0
9.75	5	15.00	0
10.00	12	15.25	0
		15.50	1
Total			196

only one zero group. *Intervals, however, should not arbitrarily be widened in order to exclude all zero classes,* because by so doing much detail may be sacrificed.

Sometimes, in order to give detail and yet avoid too numerous zero intervals, or to cover identification of data, the intervals within a table may be made of unequal widths. Tables so constructed, unless the change in interval is indicated in some striking way, are apt to be misinterpreted, because many readers notice the first interval and at once assume that the intervals following are of the same width. Table XXIII could be wrongly interpreted if the double spacing did not call attention to the fact that the width of the interval had changed.

TABLE XXIII.—FEMALE EMPLOYEES HIRED BY THE WEEK, FIRST QUARTER 1920

Source: Employment, Hours and Earnings in Prosperity and Depression (page 129)

Pay in dollars per week	Number of female employees
$ 1–$ 1.99	1
2– 2.99	27
3– 3.99	21
4– 4.99	43
5– 5.99	83
6– 6.99	39
7– 7.99	47
8– 8.99	77
9– 9.99	22
10– 11.99	95
12– 13.99	46
14– 15.99	32
16– 19.99	45
20– 23.99	23
24– 27.99	24
28– 31.99	1
32– 35.99	1
Total.....	627

By permission of National Bureau of Economic Research.

In glancing down the frequency column of Table XXIII there seems to be no marked place of concentration; a frequency of 83 occurs in the \$5 to \$5.99 class, and one of 93 in the \$10 to \$11.99 interval. On either side of these two points of concentration the frequencies diminish irregularly—group \$14 to \$15.99 has a frequency of 32, while the next above (in size) increases its frequency to 45. Unless the reader's attention is called to the unequal intervals by the change in spacing, he may conclude that the wages for women have no definite point of concentration, the number being about equally divided between the high wage, \$10 to \$11.99, and the low, \$5 to \$5.99. This, of course, is an unwarranted assumption. The frequency 83 is, really, the definite point of concentration for the series, because the larger frequency 95 was obtained by adding the number of employees having a spread in wages (\$10 to \$11.99) of twice that of the group having a frequency of 83 (\$5 to \$5.99). Likewise the 45 group has three times the interval. Attacking the problem in a different way by reducing all frequencies to dollar intervals, frequency 95 would become 47, while that of 45 reduces to 15, ($45\frac{2}{3}$). A table with unequal intervals should have the inequalities designated by means of double spacing between the unequal groups, as in Table XXIII, or by lines drawn between such groups (see Table XXI). However, in that misinterpretation may arise owing to inequality of intervals the rule may be given that: *Whenever practicable, the class intervals of a frequency table should be of equal width.* The importance of this rule will be more fully realized when the analysis of the series is taken up.

It is sometimes advocated that series with unequal intervals may, by interpolation, be broken up into classes of equal width; this is decidedly a risky practice, because if the series is discrete there is no justification for assuming that the frequency distributes itself regularly over any interval. This device therefore applies only to continuous series, and even then must be used with great care.

When it is stated that intervals "wherever practicable" should be equal it does not preclude the subdivision of a larger interval into smaller divisions; this is necessary in mortality and morbidity statistics. Deaths are more frequent per 1,000 between the age of birth to five years than from five to ten, or any multiple of 5, until the age approximates the century mark; likewise the young are more susceptible to certain diseases than the more mature. Therefore, the interval for such tables

may be five years; the first group, however, may be broken up into its component parts, as shown in Table XXIV.

TABLE XXIV.—NUMBER OF DEATHS FROM ALL CAUSES IN THE UNITED STATES, 1921

Source: Statistical Abstract, 1922 (page 76)

Age in years	Number of deaths
Under 1	160,011
1	28,907
2	13,895
3	9,923
4	7,952
Under 5	220,688
5– 9	26,058
10–14	13,205
15–19	25,989
20–24	34,937
25–29	37,040
30–34	36,654
35–39	42,281
40–44	42,628
45–49	48,371
50–54	54,813
55–59	58,057
60–64	71,240
65–69	73,785
70–74	76,621
75–79	70,598
80–84	51,409
85–89	28,398
90–94	9,778
95–99	2,371
100 and over	980
Unknown	1,178
Total	1,032,009

Table XXV also illustrates the importance of breaking up the intervals in order to portray important facts. Hours of labor, naturally, tend to concentrate on certain numbers and these have been placed between class intervals. It should be noticed that the number of workers attached to the forty-eight-hour group is not included in either the class above or below.

TABLE XXV.—WAGE EARNERS IN MANUFACTURES IN THE UNITED STATES ACCORDING TO HOURS WORKED PER WEEK, YEAR 1919

Source: Statistical Abstract, 1922 (page 283)

Hours of labor	Number employed	Per cent
44 hours and under	1,111,107	12.2
Between 44–48	346,179	3.8
48 hours	2,961,407	32.6
Between 48–54	1,496,177	16.4
54 hours	828,353	9.1
Between 54–60	1,248,854	13.7
60 hours	827,745	9.1
Over 69	276,550	3.0
Total	9,096,372	100.0

Summarizing the discussion of the class interval, it may be said that there should be between fifteen and twenty-five intervals in a table. The number, however, should be such that the width of the intervals does not distort the distribution and yet at the same time masses the material so that it may be shown in bold relief with equal intervals and few classes of a zero frequency. E. Davenport states:

> We may lay it down as a general rule that the classes should be only just broad enough to make the distribution fairly smooth, that is, there should be no vacant classes except very near the extremes of the range, and a gradual increase from one extreme up to a maximum and then a general decrease to the other extreme, if there is only one maximum in the distribution, as is, in general, the case with these (measurement of corn) populations.[1]

The data of a frequency series, however, cannot be thrown into tabular form by a mere decision as to the width of the interval. An equally important problem is the determination of the start and finish of an interval. This is the problem of the class limits.

51. The Class Limits.—In realizing that a frequency table, by virtue of its intervals, sacrifices detail, it must be apparent that the midpoint of the intervals should be representative of the

[1] DAVENPORT, E. and RIETZ, H., "Type and Variability in Corn," *Univ. Ill. Agr. Exp. Sta., Bull.* 119, p. 27.

data contained within the respective groups. It should be the normal point of distribution for the group frequency. Naturally, the point of departure of an interval has considerable to do with the value of the midpoint. Monthly salaries may be classified into $10 intervals; however, with such intervals the limits of a class may be $122.50 to $142.50, $120 to $130, or $125 to $135; the midpoint is then $127.50, $125, and $130, respectively. Monthly salaries are rarely quoted in cents. Therefore, the first grouping with a midpoint of $127.50 would hardly be representative; $125 might be chosen as a representative point of concentration, but because monthly rates are usually expressed in multiples of 10, the third grouping with a midpoint of $130 is the better one. Ten, being a point of concentration, should be the multiple used in designating the midpoints, in order to give a normal distribution to the group, because if wages concentrate on the tens, as $120 and $130, these will be the items in any grouping which would have the greatest frequency—therefore, they are more representative of the group than any other value. Age studies show a clustering around the ten-year periods 20, 30, etc. These points should be equidistant from the respective limits. *When data concentrate on certain values the class limits should be so arranged that these values are the midpoints of the classes.*

For facilitating graphic presentation and mathematical analysis, the midpoint should preferably be an integer. This will be more apparent in subsequent chapters.

The class limits should be rigidly defined and indicated on the table in order to facilitate the classification of data in the construction of the table, and the analysis of the series. If the limits are not defined, difficulties may arise in tabulating the data into the various groups. If a table should read as follows:

$ 5–$10	15
10– 15	20
15– 20	41

the question may arise as to where a $10 item should be recorded. The question cannot be answered with surety, although common sense would indicate it to belong to the second group. However, $10 is included twice in the tabular form, being the upper limit of the first group and the lower of the second group. One way to indicate the limits would be to write the above table as

$ 5 and under $10	15
10 and under 15	20

This at once gives accuracy to the limits; data between $5 and $10, not including $10, would be placed in the first group whereas the item in question would be thrown in the second group.

Another way of indicating the same material would be

$ 5–$ 9.99
10– 10.99

In this method the limits are definitely placed.

A method used to a considerable extent is that illustrated in Table XXVI. From this table it is apparent that any datum having fractional values of half or over would be recorded as of the succeeding integer—9.5 would be recorded as 10 and placed in the 10 to 14 group.

One of the best ways of designating the limits of intervals is to state under the title of the table the unit of measurement in which the data were classified; for example, age to the nearest year. The table may then be written

10–15
15–20]

In that age was recorded to the nearest year, anyone within six months of a birthday would be recorded as of the age at that birthday; as in arithmetic, when told to show only integers, 1.6 would be tabulated as 2, whereas 1.4 would be considered as 1. By this method, instead of recording the limits in fractions, as would often be necessary, as 9.5 and under 14.5, etc., they may be recorded as 10 to 15, 15 to 20, etc., the actual limits being found by deducting, from the tabulated lower limit 10, one-half of the unit of measurement. This process gives the lower limit which, by the addition of the class interval, would give the upper limit. Therefore, when data are measured to the nearest year the limits are 9.5 to 14.5; if the series is a wage series and is measured to the nearest 50 cents, the lower limit of the 10 group would be 9.75; if the series is a natural phenomena series and is measured to the nearest sixteenth of an inch, but tabulated in 2-inch intervals, the limits for the 10 to 12-inch class would be $9\frac{31}{32}$ and $11\frac{31}{32}$. Therefore, when series do not show the actual limits of their classes but record the significant unit of measurement *the limits are obtained by deducting one-half of the unit of measurement from the tabulated limits.*

TABLE XXVI.—AGE DISTRIBUTION OF RECIPIENTS OF RELIEF AND OF THE GENERAL POPULATION[1]

Age	1915				
	Number in receipt of relief	Per cent of total number in receipt of relief	General population	Per cent of general population	Per cent of general population in receipt of relief
0– 4	13,524	11.91	363,297	9.84	3.72
5– 9	18,563	15.35	335,855	9.10	5.53
10–14	16,786	13.88	302,779	8.20	5.54
15–19	9,960	8.23	314,514	8.52	3.17
20–24	5,428	4.49	359,272	9.73	1.51
25–29	5,614	4.64	346,076	9.37	1.62
30–34	6,664	5.51	299,217	8.10	2.23
35–39	7,672	6.34	289,729	7.84	2.65
40–44	6,963	6.35	252,936	6.85	2.75
45–49	5,577	4.61	217,510	5.89	2.56
50–54	4,425	3.66	182,789	4.94	2.42
55–59	4.118	3.41	134,658	3.65	3.06
60–64	3.733	3.09	104,707	2.83	3.57
65–79	3.340	2.76	76,268	2.06	4.38
70–74	2,619	2.17	54,374	1.47	4.82
75–69	1,840	1.52	33,035	0.89	5.57
80–84	1,075	0.89	16,635	0.45	6.47
85–89	435	0.36	6,601	0.18	6.59
90–94	135	0.11	1,721	0.05	7.84
95–99	36	0.03	350	0.01	10.29
100–	10	0.01	46	0.001	21.79
Unknown.....	2,388	1.98	941	0.03	
Total.....	120,905	100.00	3,693,310	100.000	

[1] HOWLAND, K. E. Am. Stat. Pub. December, 1922.

Another way frequently used in tabulation, which conserves space and yet yields accuracy, is to indicate the interval by the midpoint of the class. Table XXVI could be recorded

AGE	FREQUENCY
7	13,524
12	18,563

The limits are easily found by adding to and subtracting from the recorded midpoint one-half of the class interval.

No matter how a table is constructed it should in some manner *indicate clearly* to the reader its class limits. If time and space permit, it is perhaps the best practice to write in the actual limits in tabulation, this avoids all difficulty in interpretation.

With a determination of the class interval and limits, data may now be thrown into tabular form and given rise to a frequency table. By referring to Tables XVI and XVII the method of throwing the material into the right classes is apparent. The rigid definition of the limits should avoid all difficulties in allocating the data to the proper classes.

One more word should be said about class limits—they should as far as possible be determinate. Such headings as "under 10" and "100 and over" in a table have no determinate midpoint (see Table XVII) and these cannot be used in mathematical analysis. It is true these headings cannot always be avoided, but they should be reduced to a minimum. Another word of caution —see that the limits of the sequence of intervals do not overlap or exclude certain values.

What has been said elsewhere about the necessity of titles for tables applies equally to frequency tables.

From the above discussion it is seen that the problems of classification and tabulation of a frequency series differ greatly from those of other series, and such series require considerable care and accuracy in construction.

References

Davies, G. R.: "An Introduction to Economic Statistics," pp. 5–19.
King, W. I.: "Elements of Statistical Method," pp. 97–108.
Secrist, H.: "An Introduction to Statistical Methods."
West, Carl: "Introduction to Mathematical Statistics," pp. 24–26.
Yule, G. U.: "An Introduction to the Theory of Statistics," pp. 75–83.
Zizek, Franz: "Statistical Averages," pp. 80–91.

Problems

1. Throw four dice fifty times and record the result of each throw.

2. From the tally sheet of the above problem construct an array.

3. Construct a frequency table, with a class interval of one, of the results of the throwing of the dice.

4. Does the distribution in problem 3 agree with the distribution expected?

5. Would the distribution have been in the same proportion if there had been but five throws instead of fifty? Explain.

6. What law do the different results from different sampling illustrate?

7. Measure, with considerable refinement of measurement, the diameter of fifty nuts.

8. Construct an array from the data in problem 7.

9. *a.* What two types of series are illustrated by problems 2 and 8?

b. Wherein do they differ and what are the characteristics of each?

10. Tabulate into a frequency table having a class interval of 50 cents the data given below, thus making three tables which record the frequency distributions of the highest, lowest, and average wage.

DAILY WAGE RATES IN SPECIFIED OCCUPATIONS IN VIRGINIA, 1923–1924
Source: Monthly Labor Review, April, 1925

Occupation and race or sex	Number of employees	Daily wage rates		
		Highest	Lowest	Average
Building trades				
General contracting:				
Bricklayers:				
White	272	$15.00	$5.00	$10.14
Colored	58	11.25	6.00	8.75
Carpenters:				
White	1,336	9.00	4.00	5.82
Colored	70	6.00	2.40	3.49
Cement workers:				
White	79	10.00	3.50	6.19
Colored	50	8.00	3.50	5.03
Electricians:				
White	91	11.00	5.20	6.70
Hod carriers:				
White	44	6.00	2.00	3.68
Colored	332	5.20	2.50	4.00
Laborers:				
White	895	4.50	3.00	3.03
Colored	603	4.50	2.00	3.11
Lathers:				
White	33	10.00	4.00	5.79
Colored	27	8.00	3.15	5.39
Plasterers:				
White	93	15.00	7.20	9.61
Colored	83	10.00	6.48	8.26
Stone masons:				
White	48	10.00	5.00	8.70
Painting and paper hanging:				
Painters and decorators:				
White	233	10.00	2.00	5.05
Colored	14	5.00	3.00	3.71
Plumbing and steam fitting:				
Plumbers and gas fitters:				
White	245	10.00	3.00	6.06
Colored	6	6.00	3.00	3.60
Sheet-metal workers:				
White	169	7.00	4.00	5.82
Colored	4	3.00	3.00	3.00
Steam fitters:				
White	100	9.00	4.00	6.75

DAILEY WAGE RATES IN SPECIFIED OCCUPATIONS IN VIRGINIA, 1923–1924
(*Continued*)

Occupation and race or sex	Number of employees	Daily wage rates		
		Highest	Lowest	Average
Iron and machinery				
Blacksmiths	111	7.20	3.00	5.45
Blacksmiths' helpers	132	5.70	2.50	3.48
Boiler makers	159	7.00	2.50	5.76
Carpenters and coopers	228	6.75	3.00	4.91
Draughtsmen	143	12.00	5.00	7.24
Electricians	128	$8.00	$4.75	$5.80
Firemen	123	7.80	2.75	4.40
General help—skilled	2,580	6.75	2.70	4.19
General help—unskilled	2,927	4.00	2.13	3.01
Machinists	868	8.00	2.70	5.70
Molders	394	7.92	3.60	5.77
Pattern makers	94	9.00	4.80	6.41
Pipe fitters	61	6.24	4.95	5.40
Puddlers	32	9.84	4.95	4.93
Puddlers' helpers	17	5.16	2.70	3.03
Riveters and calkers	76	6.24	3.20	5.08
Rollers	11	10.00	4.95	8.18
Shearmen	52	4.95	2.75	3.77
Printing and engraving				
Binders:				
Males	35	9.16	3.00	6.30
Females	30	3.50	2.50	2.78
Compositors:				
Males	264	10.90	2.50	6.00
Females	7	10.00	2.50	5.40
Cutters, males	41	7.00	2.50	2.43
Engravers, males	38	10.00	3.33	7.52
Feeders:				
Males	91	5.00	1.16	2.95
Females	65	4.20	2.00	2.42
Folders:				
Males	10	7.00	1.00	3.37
Females	57	3.33	1.16	2.30
Linotype operators:				
Males	166	10.00	3.50	6.60
Females	17	7.00	2.08	4.21
Monotype casters, males	25	10.00	5.00	8.10
Monotype operators, females	12	5.42	2.83	4.08
Pressmen:				
Males	215	9.00	2.00	
Females	5	6.34	2.50	4.10
Pressmen's helpers:				
Males	75	5.25	1.50	3.04
Females	7	3.00	1.67	2.23
Proofreaders:				
Males	20	9.00	2.50	5.50
Females	33	5.60	2.00	3.37
Stereotypers, males	31	7.50	3.50	6.06

11. What other class intervals would you suggest for the above problem, and why?

12. What are the lower limits and midpoints of your classes and why do you choose such values?

13. What are the rules for the determination of class intervals and limits?

14. By glancing at the frequency distribution in problem 10, in what way would you consider them to differ from a normal distribution?

CHAPTER VIII

GRAPHIC PRESENTATION OF FREQUENCY SERIES

52. Introduction.—A frequency table, like any other table, hides somewhat in its mass of figures the relations existing between the various data, and to the layman the frequency series seems more complex because of the class interval. Once more, resort is made to graphic methods as an aid in the presentation of data.

With an understanding of the problems of the class interval and its limits, together with the knowledge of the differences in continuous and discrete series, most of the facts pertaining to graphic presentation are already known.

Naturally, the frequencies of a table should be represented on the ordinate axis, while the variable, or class intervals, should be recorded on the abscissa. This, of course, is the reverse of the procedure in the construction of an array. In the array, frequencies are not considered as the important factor.

The scales for the graph are determined in a manner similar to that described in Chap. VI; that is, they should give proportions to the chart and indicate the relationships in the right perspective, neither crowding the field of the graph nor the page itself. The scale under the abscissa axis, however, should record the *midpoint* of the class. It has been suggested also with regard to frequency series that both the vertical zero and the horizontal should be included, in order to facilitate comparison and indicate dispersion. This, however, has not yet been deemed a standard in construction.

53. Frequency Polygon or Curve.—If the intervals are of equal width, the plotting of the data becomes almost mechanical, the ordinate measurements being made from the midpoints of the classes as scaled under the abscissa. Dots are placed at the intersection of the respective abscissae and ordinates, as indicated in the table. The dots or points are usually joined by lines, giving rise to a frequency polygon or curve. A frequency polygon is the graphic presentation of a series by means of which the class frequencies are represented on the ordinate and the class interval

on the abscissa axis, the points representing the data being joined by straight lines. Polygons may be called "curves." The method of presentation is illustrated by Chart 33, which is constructed from the data in Table XVII.

By consulting Table XVII it should be noticed there are two classes, "Under $7," and "$17 and over," which are indeterminate and therefore incapable of graphic presentation. It is impossible

WEEKLY WAGES OF WOMEN MESSENGERS AND BUNDLERS IN MASSACHUSETTS STORES (RETAIL)

Source: Seventh Annual Report Minimum Wage Commission Massachusetts

CHART 33.

to assume any definite midpoint for these groups, since the group "$17 and over" may have a midvalue anywhere from $17.01 to infinity. Likewise, it is possible in some trades that premiums are paid instead of wages received; wages, therefore, may be negative. There is no justification in assuming any midpoint for the group, "Under $7." In order to draw the attention of the

chart reader to the fact that there are indeterminate groups, the end or ends of the indeterminate group should be left open. Chart 33 has both ends open. It is also good policy to indicate the frequency of indeterminate groups, when space permits, at the top of and within the field of the chart. This is a more conspicuous place than that held by regular data, which are recorded immediately above the field and over the respective midpoints.

54. Discrete Series.—The points representing data are *usually joined,* forming a polygon. Discrete series, however, are those in which the measurements increase in jumps, and the frequency of any one class does not necessarily distribute itself regularly

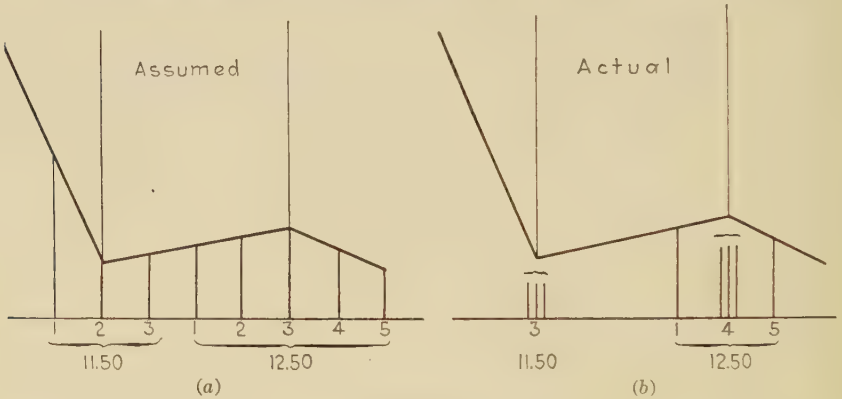


CHART 34.

over the interval. It becomes apparent, therefore, that for such series there is no justification for the polygon. The inclination of the curve in Chart 33 for the groups $11 to $12 and $12 to $13 is represented by the difference in the respective frequencies 3 and 5. This line assumes that these frequencies are distributed regularly over the two intervals, also that data could be collected for every value between the limits of these classes. This fact may be made more apparent by consulting Chart 34*a* and *b*, represent the assumed and actual distribution of the frequencies in the classes mentioned, the curve being enlarged threefold. *a* assumes an equal distribution of the frequencies over the two classes and the lines indicate that intermediate values could be obtained that is, the data may increase by small gradations. In other words, it represents the data to be continuous. *b* is the plot of the actual data as recorded in Table XVI (page 92).

This shows at once that it is highly improbable for intermediate values to be obtained, especially values expressed in odd cents. It also shows that frequencies are distributed irregularly over the intervals, but mass in groups between which there are considerable differences in values. The curve or lines, therefore, should be omitted, as it gives a false impression with regard to the class distribution and nature of the data. Therefore, for discrete series curves should only be drawn to serve as a guide in comparing the heights of the ordinates; they do not indicate the possible distribution of the data.

55. Staircase Curve.—The data of a discrete series may be represented graphically by means of a staircase curve or histo-

Weekly Wages of Women Messengers and Bundlers in Retail Stores

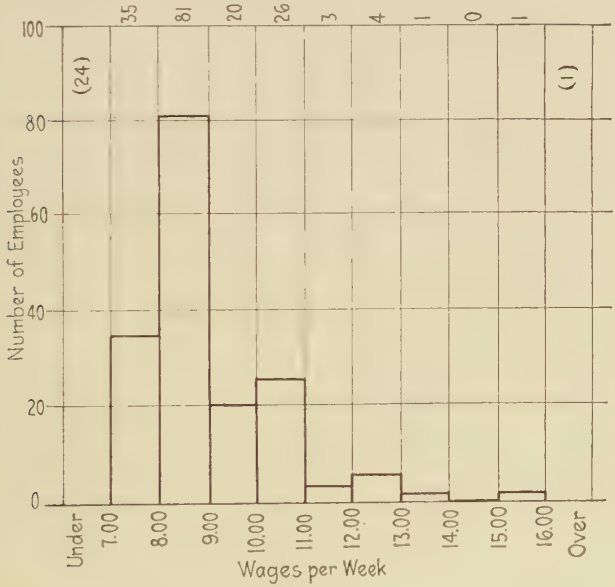


Chart 35.

gram, which presents the data in areas with no attempt to show the degrees of distribution within the respective classes. In the construction of the staircase curve the tread of the stair represents the width of the class interval; the rise, the frequency of the group. Therefore, every rectangle or stair indicates the frequency in any interval, and makes no assumption as to its distribution. The ends of the bars or stairs indicate a step-like appearance and these prevent any misinterpretation of nature of

distribution. Chart 35 is the staircase projection of the data of Chart 34. It should be noticed in the construction of Chart 35 that the figures under the abscissa are those for the lower limits of the classes and not the midpoints; furthermore, the frequencies contained within the respective classes are placed above the spaces in the field and not above the lines. This change is necessary because the frequency is distributed over the entire class and is not considered as represented by the midpoint.

56. Smoothing Curve.—Continuous series are those in which the measurements are only approximations, and, therefore, if

Comparison of the Areas of Charts 34 and 35

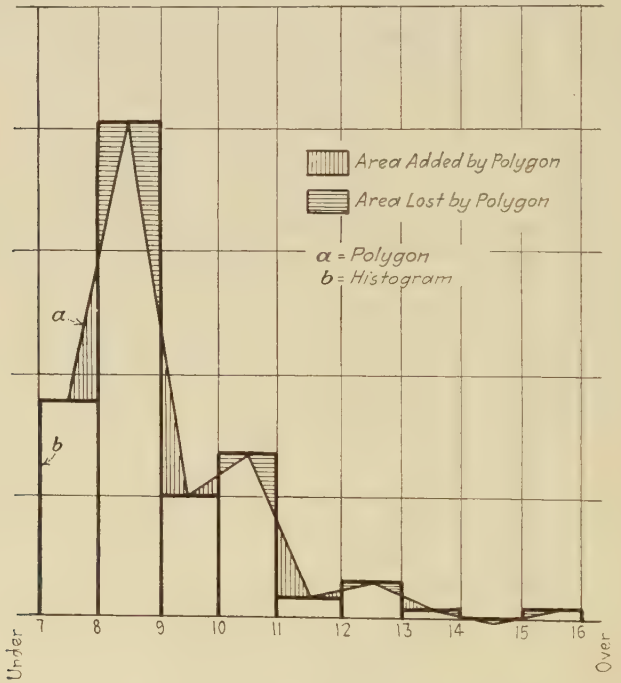


Chart 36.

sufficient data were taken and the class intervals made smaller, the ordinates, when plotted, would result in a curve. Such series may, therefore, be represented by a polygon. Since it would be possible to have all values for a variable, a curve presenting an angular appearance may be smoothed out, since the angles are the result of paucity of data. *Continuous series may then be represented by a polygon or curve, and a smoothed curve.*

Graphs at best are only approximations, and therefore smoothing introduces, unless done with greater accuracy than a graph often warrants, other errors. Smoothing may be done by free hand or by the preservation of areas. The area within a frequency curve is important in that it represents the distribution; smoothing should therefore preserve the area. The polygon, in contrast to the histogram, or staircase curve, does not give the correct area for a distribution as illustrated in Chart 36. This chart shows that the histogram and the polygon together afford excellent guides in the smoothing of a curve. It also indicates that the histogram is a better representation of the distribution, the area being larger than the polygon. However, the histogram is rarely used in graphic presentation because of the difficulty encountered in comparing series. It is hard to differentiate two histograms superimposed one upon another, whereas curves may more readily be distinguished, and understood. The histogram, however, does afford an excellent means for smoothing. A smoothed curve, unless constructed with the utmost accuracy, leads to false interpretations, just as the frequency polygon does for discrete series. Both these forms are used, but they should be used with great caution.

57. Series with Unequal Intervals.—A series in which the class intervals are unequal may demand graphic presentation. The only method, apart from the ogive—which will be discussed later—is to break up the unequal groups into distributions of equal classes. In Table XXVII, it is evident that the frequencies are affected considerably by the inequality of the class interval. The frequencies in column B cannot be plotted from the midpoint of each class to derive a frequency polygon because of the inequality in intervals. By dividing these frequencies by the multiple of the actual class interval to a standard class interval, however, the average distribution is found in terms of an equal interval. Column C indicates the various multiples of the standard, which is one week. These multiples are divided into their respective frequencies, resulting in column D, all frequencies being now expressed on the weekly basis. The data are then plotted; measurements are made from the *actual midpoint* of the classes. The range, however, in this table is so large that a very large graph would have to be resorted to if rectilinear paper were used, in order to bring out the distribution. Double logarithmic paper solves the difficulty at once and furthermore gives a more symmetrical distribution, as shown in Chart 37.

The resulting curve is almost a straight line, whereas, by consulting the frequencies from which it was constructed, it will be seen that on rectilinear paper the curve would be a J type. This fact once more emphasizes the advantages of logarithmic paper—in this example it permits data with wide range to be brought into a small compass and also gives a more symmetrical distribution.

DOUBLE LOGARITHMIC PROJECTION OF TABLE XXVII

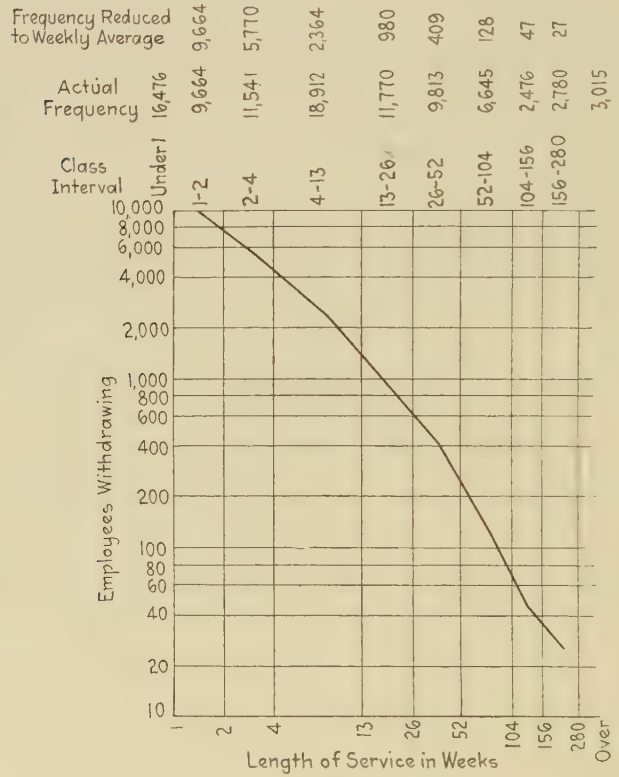


CHART 37.

In contrast to the logarithmic paper discussed in Chap. VI, it will be observed that Chart 37 has double logarithmic projection. For frequency series both axes may have logarithmic scales, although it is more usual to have the abscissa logarithmic and the ordinate arithmetic; this is the reverse of logarithmic paper for other series (compare Charts 29 and 39).

TABLE XXVII.—LENGTH OF SERVICE OF ACTIVE EMPLOYEES WITHDRAWING FROM EMPLOYMENT DURING 1917–1918

Source: Monthly Labor Review, June, 1920 (page 51)

Length of service A	Number B	Class interval C	Frequency per class interval of one week D
1 week or less	16,476	1	16,476
Over 1 week to 2 weeks	9,664	1	9,664
Over 2 weeks to 1 month	11,541	2	5,770
Over 1 month to 3 months	18,912	8	2,364
Over 3 months to 6 months	11,770	12	980
Over 6 months to 1 year	9,813	24	409
Over 1 year to 2 years	6,645	52	128
Over 2 years to 3 years	2,476	52	47
Over 3 years to 5 years	2,780	104	27
Over 5 years	3,015		
Total	93,092		

TABLE XXVIII.—DURATION OF DISABILITY, AND NUMBER OF CASES IN ALL OCCUPATIONS, YEARS 1912–1916

Source: Monthly Labor Review, March, 1920

Duration of disability A	Number B	Class interval C	Frequency per class interval of one week
Under 1 week	8,303		
1 week and under 2 weeks	12,071	1	12,071
2 weeks and under 4 weeks	10,468	2	5,234
4 weeks and under 12 weeks	9,418	8	1,177
12 weeks and under 25 weeks	2,269	13	174
25 weeks and over	956		
Total	43,485		

DURATION OF DISABILITY AND NUMBER OF CASES IN ALL OCCUPATIONS, YEARS 1912–1916

Data in Table XXIX

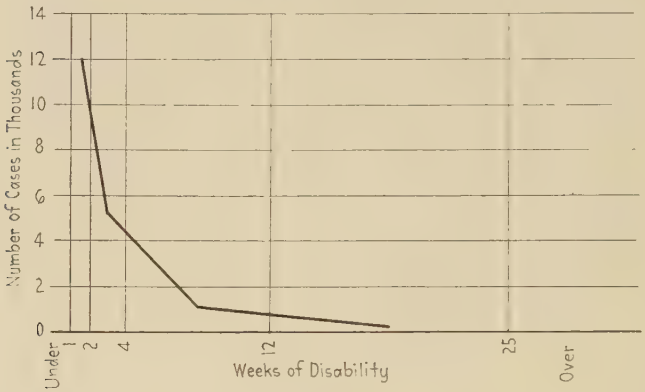


CHART 38.

LOGARITHMIC PROJECTION OF TABLE XXIX

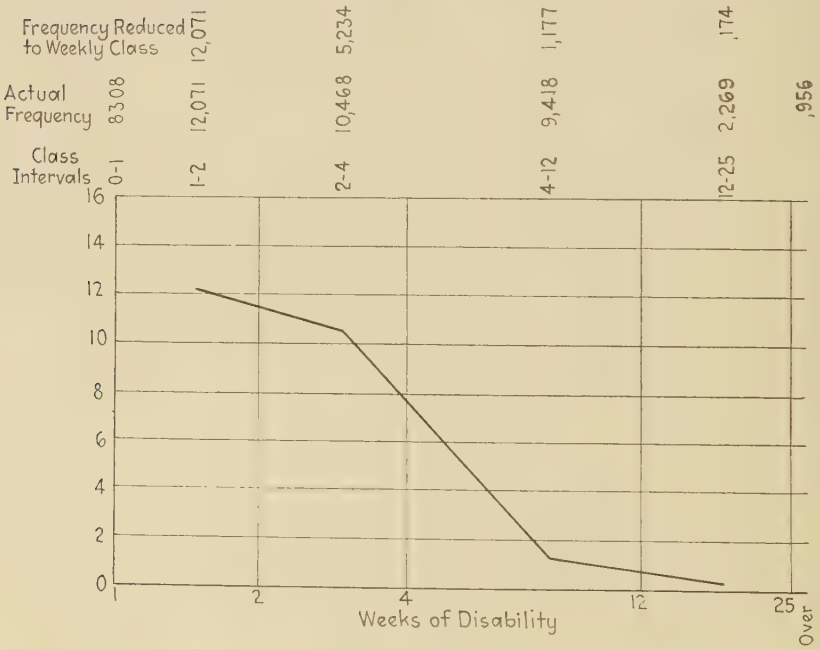


CHART 39.

An example of the rectilinear charting of Table XXVIII with unequal intervals is given in Chart 38. This chart illustrates an asymmetrical J-shaped curve, whereas when the same data are plotted on horizontal logarithmic paper the curve takes a very different conformation, as in Chart 39. The reason for such differences should be apparent when it is recalled that at the lower end of the range of a table a small difference in size gives a relatively larger ratio change than the same difference in the upper limits of the series. Thus the logarithmic scale lengthens the graphic presentation of the lower limits and contracts the upper limits of a series. Logarithmic scales are useful in most types of series, but such paper can never be used for series indicating a zero or negative quantity.

In Charts 37, 38, and 39 it should be observed that the *limits* instead of the *midpoint* of the classes are tabulated below the abscissa. This departure from the usual procedure for frequency polygons is made because with such irregularities in the intervals it would be difficult to visualize the class limits. The data, however, are plotted from the midpoint of the intervals as in ordinary tables. This type of chart is tedious in construction and has very little to commend it as an aid in presentation. Its use should therefore be reduced to a minimum, the tabular form usually being sufficient.

58. Comparison of Curves.—Very often it is desirable to superimpose one frequency curve upon another, but if they are plotted in absolute numbers the divergence in the range may prohibit bringing the curves into juxtaposition. In Table XXIX the series vary from single frequencies to millions. By expressing each frequency as a percentage of the total population for the respective series, the range of all series, no matter the divergence in absolute numbers, falls within 0 to 100 per cent. In Table XXIX the range for the four series does not exceed 15 per cent, as indicated in the percentage columns of the table and the graphic illustration of Chart 40. Chart 40 indicates the "native parentage" group to be maintaining its proportion of children when compared to other groups, also that "foreign-born white" spread out in the upper range of the series, while the Japanese tend to be in the prime of life, there being few Japanese over fifty. All these facts are hidden in the mass of figures, but the chart aids in the analysis of such figures by indicating readily the various relationships.

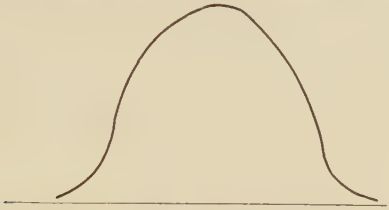
TABLE XXIX.—ACTUAL AND PERCENTAGE DISTRIBUTION ACCORDING TO AGE GROUPS OF THE TOTAL POPULATION OF THE UNITED STATES, NATIVE WHITE, FOREIGN-BORN WHITE, AND JAPANESE

Source: Fourteenth Census of the United States

Age	All races		White				Japanese	
			Native-born		Foreign-born			
	Number	Per cent	Number	Per cent	Number	Per cent	Number	Per cent
All ages	105,710,620	100.0	58,421,957	100.0	13,712,754	100.0	111,010	100.0
Under 5 years	11,573,230	10.9	7,366,530	12.6	44,984	0.3	19,029	17.1
Under 1 year	2,257,255	2.1	1,453,404	2.5	4,569	(1)	4,704	4.2
5 to 9 years	11,398,075	10.8	6,977,863	11.9	169,884	1.2	7,743	7.0
10 to 14 years	10, 41,137	10.1	6,455,709	11.1	331,362	2.4	2,569	2.3
15 to 19 years	9,430,556	8.9	5,599,046	9.6	527,942	3.9	4,342	3.9
20 to 24 years	9,277,021	8.8	5,176,707	8.9	926,844	6.8	9,636	8.7
25 to 29 years	9,086,491	8.6	4,764,802	8.2	1,454,363	10.6	11,270	10.2
30 to 34 years	8,071,193	7.6	4,096,041	7.0	1,651,475	12.0	15,253	13.7
35 to 39 years	7,775,281	7.4	2,815,852	6.5	1,737,805	12.7	15,394	13.9
40 to 44 years	6,345,557	6.0	3,099,330	5.3	1,428,099	10.4	12,762	11.5
45 to 49 years	5,763,620	5.5	2,753,013	4.7	1,299,675	9.5	7,181	6.5
50 to 54 years	4,734,873	4.5	2,236,700	3.8	1,167,377	8.5	3,515	3.2
55 to 59 years	3,549,124	3.4	1,719,190	2.9	908,722	6.6	1,214	1.1
60 to 64 years	2,982,548	2.8	1,530,129	2.6	715,731	5.2	489	0.4
65 to 69 years	2,068,475	2.0	1,126,722	1.9	519,226	3.8	151	0.1
70 to 74 years	1,395,036	1.3	784,183	1.3	381,325	2.8	31	(2)
75 to 79 years	856,560	0.8	487,797	0.8	246,487	1.8	12	(2)
80 to 84 years	402,779	0.4	228,018	0.4	118,935	0.9	2	(2)
85 to 89 years	156,539	0.1	87,530	0.1	47,473	0.3	5	(2)
90 to 94 years	39,980	(2)	20,159	(1)	11,717	0.1		
95 to 99 years	9,579	(2)	3,853	(1)	2,579	(1)		
100 years and over	4,267	(2)	614	(1)	485	(1)		
Age unknown	148,699	0.1	92,169	0.2	20,264	0.1	412	0.4

59. Types of Curves.—By comparing the curves plotted in this chapter and visualizing the conformations of the series not given graphic presentation, it should be apparent that there are differences in the degree of symmetry. Some curves slope immediately down to the right, as in Chart 38, which is a J-shaped curve, while others climb rapidly to a peak and then decline gradually, as that for foreign-born white in Chart 40, while others climb gradually and decline quickly. Studies of natural

phenomena often exhibit a perfectly symmetrical curve, called the "normal curve of error." This curve, however, is rarely found in economic studies. The normal curve is illustrated below. In such a curve the frequencies increase or decrease



in the same proportions on either side of the peak called "norm." If the curve were bisected through its apex to the

PERCENTAGE DISTRIBUTION, ACCORDING TO AGE GROUPS, OF TOTAL POPULATION OF THE UNITED STATES, NATIVE BORN WHITE, FOREIGN BORN WHITE AND JAPANESE, AS OF THE CENSUS 1920

Source: Bureau of Census

Data in Table XXIX

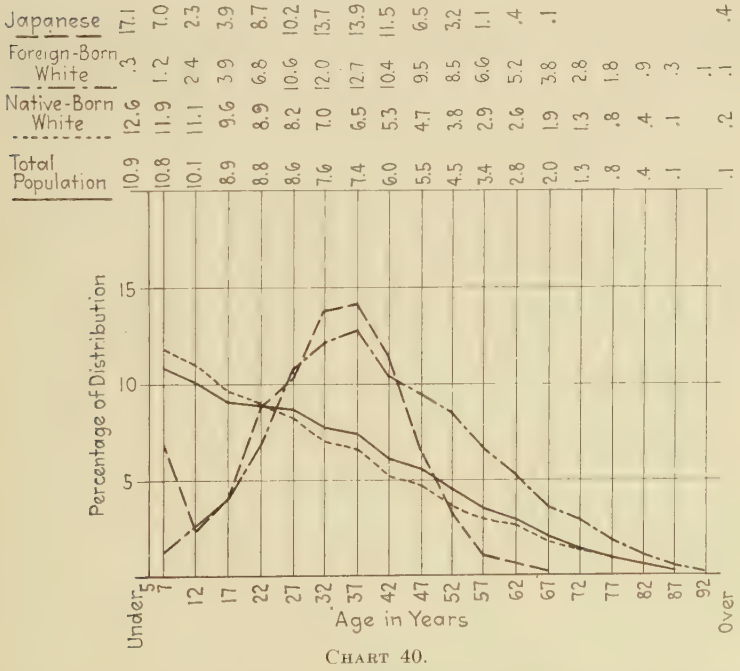


CHART 40.

base, the two halves would be identical. Such a curve indicates that the data concentrates at a definite point, called the "norm."

The curve because of its symmetry has measures of position (discussed later) which are identical. Therefore the divergence found in certain measures of position in other series forms the basis of the measurement of skewness or divergence.

Curves may be divided into two groups:

1. Symmetrical. 2. Asymmetrical.

Naturally, there are various degree of asymmetry, such as moderate asymmetry and extreme asymmetry, the extreme types being those of the J- and U-shaped curves. The importance of the direction and degree of symmetry will be taken up in the chapter on skewness.

60. The Ogive.—In using tables and graphs it is sometimes desired to know not only the class frequencies of a table but also the number of items below (less than) a particular class, or those above (more than) a definite interval. This is particularly true of wage and income studies when it is often important to know how many persons receive above or below a certain standard income. The simple frequency does not give this information at once, but it may be obtained by adding the required frequencies. To answer such questions and afford easy reference the cumulative frequency table may be constructed. In this table the frequencies are cumulated either up or down. Such a table is illustrated below.

In comparing columns C and D of Table XXX the "less than" figures read from and including the upper limit of the class, while the "more than" include and read from the lower limit of the class. That is, the group \$3 to \$3.99, with a frequency of 21, is so situated in the table that there are 49 items occurring between \$3.99 and \$1 inclusive. At the same time there are 599 items between \$3 and \$35.99.

The data contained in columns C and D may be represented graphically by curves called ogives. In that data may be cumulated either "more than" or "less than," two curves may represent the data, which, owing to the nature of the data, take different conformations. In plotting a "less than" ogive the cumulated data must be plotted from the *upper limit* of the classes, the curve ascending to the right. A "more than" ogive requires that the data be plotted at the *lower limit* of the class, the curve descending from the right. There should be no confusion with regard to the limit to use in plotting if the idea of cumulation is thoroughly understood.

TABLE XXX.—FEMALE EMPLOYEES HIRED BY THE WEEK, FIRST QUARTER, 1920

Source: "Employment, Hours, and Earnings in Prosperity and Depression" (National Bureau of Economic Research)

Pay in dollars per week	Number of female employees		
	Simple frequency	Cumulative frequency	
		"Less than"	"More than"
A	B	C	D
$ 1–$ 1.99	1	1	627
2– 2.99	27	28	626
3– 3.99	21	49	599
4– 4.99	43	92	578
5– 5.99	83	175	535
6– 6.99	39	214	452
7– 7.99	47	261	413
8– 8.99	77	338	366
9– 9.99	22	360	289
10– 11.99	95	455	267
12– 13.99	46	501	172
14– 15.99	32	533	126
16– 19.99	45	578	94
20– 23.99	23	601	49
24– 27.99	24	625	26
28– 31.99	1	626	2
32– 35.99	1	627	1
Total	627		

By permission of the National Bureau of Economic Research.

The ogive is a very useful curve, for, by the process of accumulation, tables with unequal intervals may be represented graphically without the laborious process of breaking up the frequencies into averages for a given standard, as indicated in Table XXVII. All that is required in an ogive presentation of unequal intervals is that the accumulated frequency be plotted at the *actual* upper limit (if less than) or lower limit (for more than) of the class, irre-

spective of the width of such class, the scale of the abscissa throwing the plot according to the width of the interval. The ogives for Table XXX are illustrated in Chart 41. This chart, it should be noticed, represents unequal intervals with ease.

Ogive Distribution of Data in Table XXX

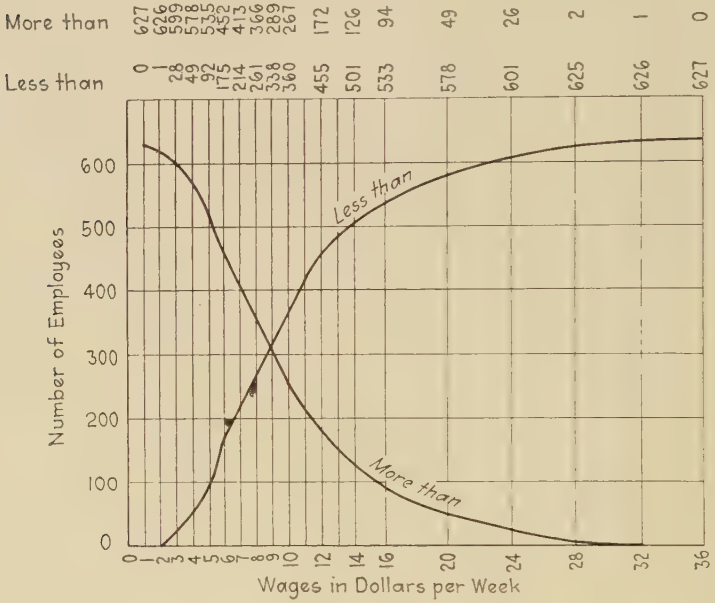


Chart 41.

61. Interpolation by Means of Ogive.—The ogive may also be used for interpolating between given values. Interpolation is the measuring of values within given values. If a series is continuous, it is possible to measure either the number of frequencies occurring at certain points within given values, or values for certain frequencies may be found. This is illustrated in Chart 42. In this chart assume that it is desired to know how large a percentage of the population of the foreign-born white is found between the ages of thirty-five and thirty-eight. By means of the perpendicular erected at X (age thirty-eight), which cuts the curve in Y, the vertical distance MY represents the accumulation of frequencies from thirty-five to thirty-eight, which, when measured on the scale at the left, gives a value of 6.8 per cent. Second, assume that a value is required for a given frequency, the frequency being 80 per cent. By measuring the distance

from a known value (55) to the point of intersection of the curve at the required frequency (80), it is found that the age is fifty-seven. When such interpolation is required it is customary to place such derived data at the right of the graph. (In this illustration the graph has been defaced in order to make the process apparent.)

OGIVE OF PERCENTAGE DISTRIBUTION OF TABLE XXIX

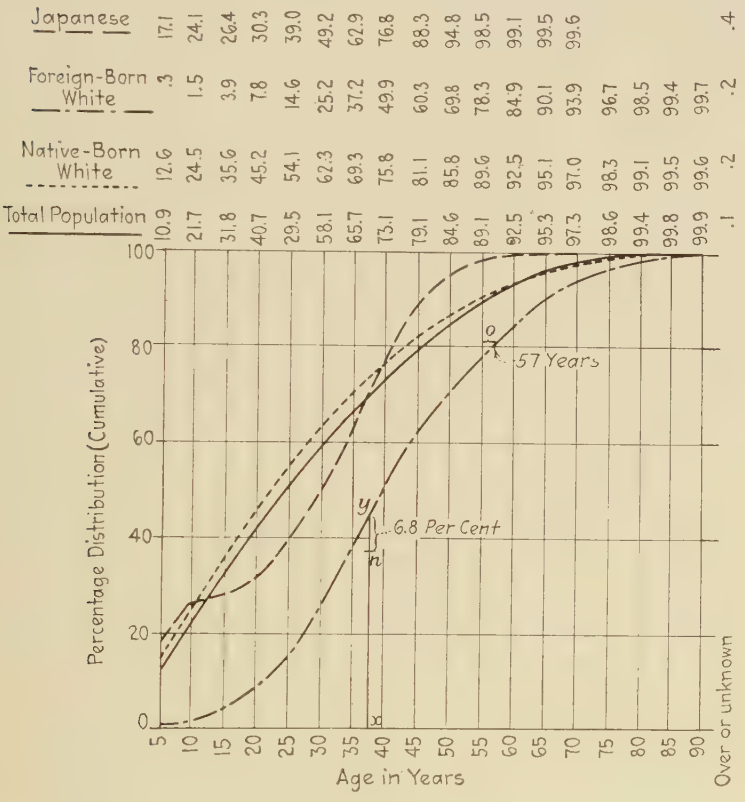


CHART 42.

Interpolation is very often used in breaking up unequal intervals into equal groupings. Such method, however, is hardly the function of a graph. It may be justified when the series are continuous, but never when discrete.

The ogive is a type of graphic presentation which permits by graphic methods the rough measurement of the median, quartiles, and percentiles. These will be discussed later. Because of the

accumulation of data the ogive is usually smoothed, even for discrete series.

In comparison of curves the ogive has a distinct advantage over the simple frequency curve, because its curve pursues but one direction, the only variance being in the degree of inclination. Ogives, therefore, when superimposed upon one another will not intersect one another to any marked extent, as is the case in the comparison of simple polygons. This will be evident by comparing Charts 40 and 42; the ogive presents a better comparative study than the frequency polygon.

Double Logarithmic Chart of the Same Data as in Chart 42

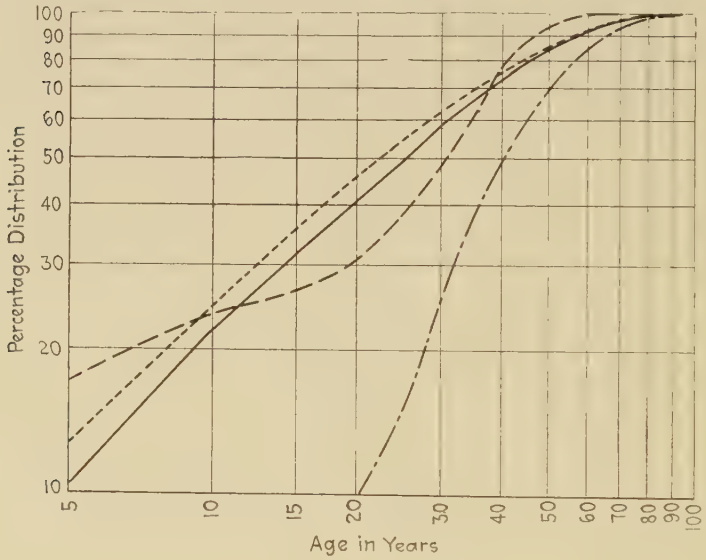


Chart 43.

Therefore, the ogive is one of the most useful types of curves for frequency series. Its construction sometimes permits the graphic presentation of classes with indeterminate limits, depending upon the location of the indeterminate class and whether it is a "more than" or "less than" ogive.

Logarithmic paper may also be used to advantage in the construction of ogives. Since the data are cumulative, it is customary to use double logarithmic projection. Chart 43 is the logarithmic presentation of Chart 42. By comparison of these two charts the smoothing effect of logarithmic paper is again

evident. The significance of the degree of inclination of the curves in such paper has been discussed elsewhere.

62. The Lorenz Curve.—Another curve used in the graphic presentation of frequency series is called the Lorenz curve. This curve permits the reader to visualize dispersion or scatteration.

Everyone is conscious of the inequality of income. It is well known that a large percentage of wealth is held in the hands of relatively few individuals. The Lorenz curve gives a pictorial presentation of distribution. If the days of absolute equality were ever ushered in, which is the desire of a very small group in society, 10 per cent of the people would possess 10 per cent of the total wealth; 20 per cent would own the same percentage of wealth—the percentage of wealth would correspond to the percentage of the population. If the data were plotted with percentages of wealth represented on the abscissa and those for population scaled on the ordinate, the points when joined would result in a straight line inclined to the base at an angle of 45 degrees. This line would be the line of equal distribution. The degree with which the actual distribution deviates from equality may, therefore, be indicated by the deviation of the curve from the line of equal distribution. For the curve of income in the United States for the year 1918, see "Income in the United States," page 141, by the National Bureau of Economic Research.

TABLE XXXI.—ACTUAL AND PERCENTAGE DISTRIBUTION OF THE ACREAGE IN FARMS AND THE VALUE OF ALL FARM PROPERTY, ACCORDING TO ACREAGE GROUPS

Source: Fourteenth Census of the United States

Acres	Acreage			Value of all property		
	Number	Per cent	Cumulative per cent	Value	Per cent	Cumulative per cent
A	B	C	D	E	F	G
Under 20	8,688,899	0.9	0.9	2,452,977,205	3.1	3.1
20– 49	48,464,330	5.1	6.0	5,864,412,204	7.5	10.6
50– 99	105,630,796	11.1	17.1	11,183,267,488	14.4	25.0
100–174	194,681,260	20.4	37.5	20,901,789,301	26.8	51.8
175–499	276,806,995	29.0	66.5	26,389,647,868	33.9	85.7
500–999	100,975,916	10.6	77.1	5,533,774,473	7.1	92.8
1,000 and over	220,635,519	23.1	100.	5,598,321,799	7.2	100.
Total.....	9,555,883,715	....	....	77,924,100,338		

As an illustration of the method involved, suppose it is desired to find out whether or not the value of all farm property increases in proportion to the acreage held. This problem requires data of the total number of acres found in the different sized farms, and the aggregate value of farm property within each acreage class. Table XXXI presents the absolute figures and the cumulated percentage data.

PERCENTAGE DISTRIBUTION OF ACREAGE IN FARMS RELATED TO THE VALUE OF FARM PROPERTY
Data in Table XXXI

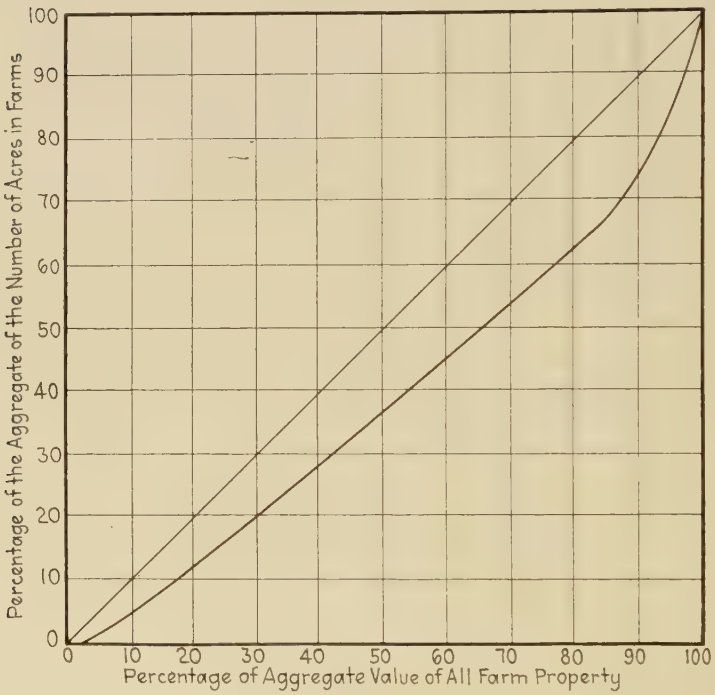


CHART 44.

The Lorenz curve is obtained by plotting the cumulative percentage figures of columns D and G. D is scaled on the ordinate, while G occupies the abscissa (Chart 44).

The curve deviates considerably from the line of proportional change in the upper range of the percentage figures, the interpretation being that the smaller farms have a greater percentage of the value of farm property than the larger farms, since, as the

percentage of farms included increases, the value lags or falls from the line of equal change.

The defect of such curves is, of course, the fact that they are made from relative numbers. The logarithmic curve may be used to obviate such defect, a double projection being used. Logarithmic projections are often used in income studies and were first advocated by Pareto. For illustrations of these see pages 121 and 129 of "Income of the United States," by the National Bureau of Economic Research.

In this chapter most of the graphic methods used in charting frequency series have been discussed. The charts differ somewhat in construction from those discussed in Chap. VI. While charts must at all time be considered as aids in analysis and not as analysis, by appealing to the eye they become effective weapons in the hands of the statistician, and therefore their construction should receive considerable attention.

References

Bowley, A. L.: "Elements of Statistics," chap. VII.

Brinton, W. C.: "Graphic Methods for Presenting Facts."

Chaddock, R. E.: "The Graphic Representation of a Frequency Distribution," *Quar. Pub. Amer. Stat. Assoc.*, June, 1921.

Clark, E.: "The Horizontal Zero in Frequency Diagrams," *Quar. Pub. Amer. Stat. Assoc.*, June, 1917.

Elderton, W. P. and E. M.: "Primer of Statistics," chap. III.

Karsten, K. G.: "Charts and Graphs."

Kelley, Truman L.: "Statistical Method," pp. 29–37.

King, W. I.: "Elements of Statistical Method," chap. XI.

Secrist, H.: "An Introduction to Statistical Methods," chap. VII.

West, Carl: "Introduction to Mathematical Statistics," chap. III.

Yule, G. U.: "An Introduction to the Theory of Statistics," pp. 83–105.

Problems

1. Plot, in the same field, on logarithmic paper, the polygons or curves for the highest, lowest, and average daily wage paid in Virginia 1923 to 1924 as found in problem 10, Chap. VII.

2. Which of the three curves plotted above is the nearest approach to a normal distribution curve? Why?

3. Make an histogram or staircase curve for the average daily wage of the same problem as above.

4. *a.* When are you justified in smoothing a polygon?

b. What method would you adopt for smoothing?

5. Draw ogives for the highest wage group above, making one "less than" and the other "more than."

6. Given the following table:

MANUFACTURES, SIZE OF ESTABLISHMENT BY VALUE OF PRODUCTS, 1919
Source: U. S. Census

Value of product	Establishments		Average number of wage earners	
	No. A	P.C. B	No. C	P.C. D
All classes	290,105	100.0	9,096,372	100.0
Less than $5,000	65,485	22.6	45,813	0.5
$ 5,000–$ 20,000	87,440	30.1	249,722	2.7
20,000– 100,000	77,911	26.9	793,528	8.7
100.000– 500,000	39,647	13.7	1,719,982	18.9
500,000– 1,000,000	9,208	3.2	1,114,615	12.3
1,000,000 and over	10,414	3.6	5,172,712	56.9

a. What methods could be used to break up the unequal class intervals into equal classes of $5,000?

b. Apply one method to the data in column C. Also plot the results on double logarithmic paper (three-cycle).

7. From the data given in the table above construct a Lorenz curve.

8. What is the interpretation of the results of the Lorenz curve of problem 7?

CHAPTER IX

METHODS OF ANALYZING A FREQUENCY SERIES

63. Introduction.—In discussing the array it was found that a mere recitation or presentation of the data arranged in successive order according to magnitude aided little in analysis—the detail obscured the analysis. The frequency table was devised in order to sacrifice some of the detail, to compress the data into less space, and so obtain perspicuity. No matter how small a table may be, however, it is very hard to characterize a series by listening to, or reading, its contents. Graphic methods aid in presentation, but such methods, as already indicated, are only approximations. Usually a problem requires some means by which a series may be characterized in order to facilitate the comparison within or between series.

64. Types of Analysis.—A series may be analyzed or characterized in three ways:

1. By measuring the "central tendency" of the data, that is, their concentration around a norm.

2. By measuring the degree of concentration around the norm, that is, the degree of divergence from type.

3. By measuring the direction of concentration around the norm.

In the graphic presentation of Table XXXII as shown in Chart 45, these three concepts are illustrated.

In this chart is should be observed;

1. There is a point of maximum frequency in both series. The groups 55 to 59 and 65 to 69 are the norms for the laborers' and the machinists' wage series, respectively.

2. There is considerable divergence in the wages of machinists, since the frequencies on either side of the norm do not fall off rapidly; the polygon has a flattened appearance in contrast to that for laborers' wages. The latter series reaches its peak in the group, 55 to 59. The frequencies, however, on either side fall off rapidly, indicating that, generally, laborers' wages do not vary to any great extent.

TABLE XXXII.—WAGES PER HOUR OF MACHINISTS AND LABORERS IN FACTORY X

Rate per hour (cents)	Number	
	Machinists	Laborers
Total	96	107
49–54	2	21
55–59	7	40
60–64	14	30
65–69	25	7
70–74	23	6
75–79	15	2
80–84	10	1

CHART 45.

3. The conformations of the two polygons differ. That for the laborers' series rises sharply, at the lower wage rates, to a peak, then falls rapidly and finally tapers off gradually. This polygon has what is termed a "positive skew." The machinist polygon, on the other hand, rises less abruptly to the norm and finally declines sharply.

These observations indicate that a series may differ in three respects: the first, that of central tendency, is called "averages," the second, the degree of variation from the norm, is termed "dispersion;" while the third, the direction of divergence, is called "skewness."

A series may be analyzed in any or all of these ways, but from the above illustration it should be evident that no *one* measure indicates all the characteristics of a series. Therefore, conclusions or analysis should not be derived by hurried resort to any one method. It is essential to remember that any one method sacrifices detail and gives a view, as it were, from some particular vantage point, other, but different, views being possible. Two series may have identical measures of position (averages) or dispersion and yet differ considerable in other respects.

65. Types of Averages.—An average is a measure of position which expresses the tendency of data to concentrate in various ways at definite measurements of variable phenomena.

The most common measures of position are the arithmetic mean, median, and mode. Others used less frequently are the quartiles, percentiles, harmonic mean, and geometric mean.

The arithmetic mean of a series is the measurement of that variable which divides the data in such a way that the sum of the deviations (in units of measurement) of all data from the mean equals zero. In other words, it is the result obtained by dividing the sum of the measurements of all the variables by their number.

In Table XXXIII, with Chart 46 attached, it is seen that the combined weights of the seven persons totaled 1,155 pounds. The average weight, therefore, is $\frac{1,155}{7}$, or 165 pounds. The chart and the table illustrate how the sum of the deviations from 165 pounds is zero.

The median is that position on the scale, whether measuring quantities or qualities, which has as many items above it as below it, when the items are arranged according to magnitude of

TABLE XXXIII.—TABLE SHOWING THE SUM OF THE DEVIATIONS FROM THE MEAN EQUALS ZERO

Weights	Deviations from mean	
	+	−
Total 1,155	35	35
154	..	11
154	..	11
159	..	6
160	..	5
163	..	2
165	Mean	Mean
200	35	

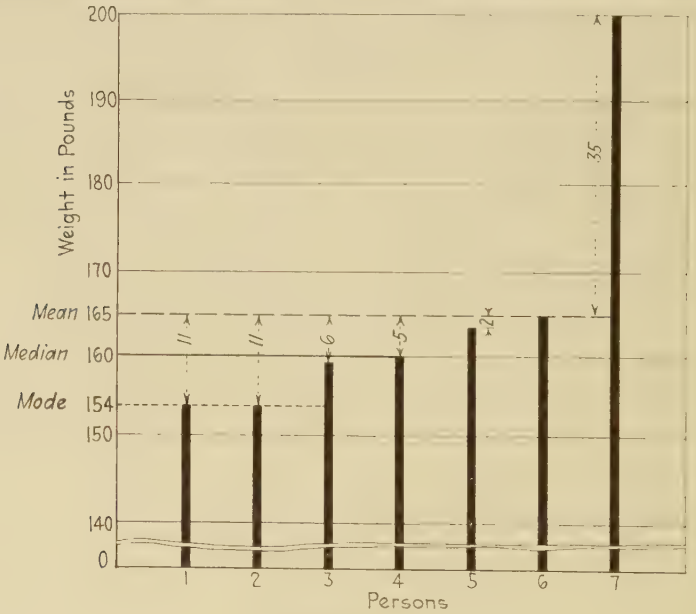


CHART 46.

size or quality. In the above illustration the weights of seven persons were recorded. The fourth item, therefore, divides the total frequency seven into two equal parts. This item is the

median item, the measurement of which gives the median, 160 pounds.

The mode of a series is the measurement of the item of greatest frequency. In the illustration the weight of 154 pounds occurs twice in contrast to the other weights; 154, therefore, is the modal weight.

66. The Desired Attributes of an Average.—These three averages differ from one another in certain respects, as indicated by the lines drawn for the respective measures in Chart 46. These differences will be discussed as each type of average is discussed in detail. The quartiles, percentiles, geometric means and harmonic mean will also be defined later. However, all averages should possess certain qualities.

(*a*) In the first place, it goes without saying that an average should be rigidly defined, and not left to the mere estimation of the observer. An average that was merely estimated would depend too largely on the observer as well as on the data. (*b*) An average should be based on all the observations made; if not, it is not really characteristic of the whole distribution. (*c*) It is desirable that the average should possess some simple and obvious characteristic properties to render its general nature readily comprehensible; an average should not be of too abstract a mathematical character. (*d*) It is, of course, desirable that an average should be calculated with reasonable ease and rapidity. Other things being equal, the easier calculated is the better of the two forms of average. (*e*) It is desirable that the average should be as little affected as may be possible by what we have termed fluctuations of sampling. If different samples be drawn from the same material, however carefully they may be taken, the average of the samples will rarely be quite the same, but one form of average may show greater differences than another . . . (*f*) Finally, by far the most important desideratum is, that the measure chosen shall lend itself to algebraic treatment.[1]

No one average possesses all these attributes and the choice of the average depends largely upon the purpose of the problem. Purpose dominates the analysis; the average should then be chosen which would give the maximum number of the above attributes.

References

KELLEY, TRUMAN L.: "Statistical Method," pp. 44–45.
YULE, G. U.: "An Introduction to the Theory of Statistics," pp. 106–108.

[1] YULE, G. U., "An Introduction to the Theory of Statistics," p. 108.

CHAPTER X

THE MEAN

I. THE ARITHMETIC MEAN

67. Introduction.—The arithmetic mean is one of the averages most familiar to the layman, who generally terms it the average. In this text, however, as there are different kinds of averages it will be called the mean. If occasion arises, to avoid confusion with the geometric or harmonic mean, it will be given its full designation, "arithmetic mean." The arithmetic mean has already been defined as that point in a series from which the sum of the deviations of measurement equals zero. It is now necessary to formulate methods for its calculation.

68. The Simple Method of Computation.—In a simple problem, such as that in Table XXXIII, or in the more complex Table XXXIV, the so-called simple method may be characterized by the following formula;

$$\text{Mean} = \frac{\Sigma (f \times v)}{N}.$$

In this formula, and all others in this text, the Greek letter Σ (sigma) represents the mathematical concept of summation, while N signifies the total frequency or number of variables measured in a table. V in this particular formula indicates the measurements of the variable. The entire formula, therefore, represents the mean to equal the aggregate of the measurements of all the variables divided by their number.

Table XXXIV lists the data in class intervals. The interval seemingly complicates, in contrast to the simple Table XXXIII, the derivation of the mean. The difficulty, however, is more apparent than real, for it should be recalled from the discussion of classification and tabulation that the midpoint of a class interval should be representative of the group. Therefore, the class \$15 to \$20 of Table XXXIV, which has a frequency of 10, should be considered as comprising ten persons, each receiving a \$17 weekly wage, since \$17 is the mean point of the class then $f \times$

V would equal $17 \times 10 = 170$. By continuing the process of multiplying the midpoint of the classes by their respective frequencies (column C) the numerator of the formula may be derived by summating their products. This when divided by the total frequency (103), yields a mean of $30.79.

If the table had been one expressing weekly wage *rates*, which is a discrete series—since such rates are rarely paid in odd cents—the mean would be given as $30.75, the nearest known gradation to the computed mean. *The mean should never indicate greater accuracy in measurement than the accuracy with which the data were collected.* In Table XXXIV, however, weekly wages (not rates) differ by small gradations and it is possible for $30.79 to be an actual value.

The above method of obtaining the mean is a laborious one and especially liable to error if the frequencies and range are large. These defects may be remedied by the application of the interval method.

TABLE XXXIV.—METHODS OF THE CALCULATION OF THE MEAN WAGES PER WEEK IN FACTORY X MEASURED TO NEAREST DOLLAR

Wages	f	Long method		Interval method		
		Mid-point	Product $f \times$ mid	Dev. in intervals	$d \times f$	Product
A	B	C	D	E	F	G
$15–$20	10	17	170	−3	−3 × 10	−30
20– 25	19	22	418	−2	−2 × 19	−38
25– 30	20	27	540	−1	−1 × 20	−20
						−88
30– 35	21	32	672	0	0	0
35– 40	18	37	666	+1	1 × 18	18
40– 45	0	42	0	+2	2 × 0	0
45– 50	15	47	705	+3	3 × 15	45
						+63
Total.....	103	..	3,171	...	.....	−25

Long method.

$$\text{Mean} = \frac{\Sigma(f \times v)}{N}$$
$$= \frac{3{,}171}{103} = \underline{\underline{\$30.79}}$$

Interval method.

$$\text{Mean} = A + \frac{\Sigma(d \times f)c}{N}$$
$$= 32 + \frac{(-25)5}{103}$$
$$= 32 - 1.213 = \underline{\underline{\$30.79}}$$

69. The Interval Method.—The interval method of calculating the mean first assumes some one interval, which may be any, to be the arbitrary origin, the midpoint of which is the *assumed mean.* In Table XXXIV, the group \$30 to \$35 is arbitrarily chosen as the point of origin. From the chapter on the frequency table it will be recalled that a class interval together with its limits should be so arranged that the midpoint of the class should be representative of all variables within the group. Therefore, in Table XXXIV, the arbitrary or *assumed mean* (A) is 32, the midpoint of the class \$30 to \$35.

If 32 were the actual mean, by definition the sum of the deviation from it would equal zero, so that

$$\text{Mean} = A.$$

To prove that Mean = A it is necessary to measure the deviations from the arbitrary origin. In this method deviations are *expressed in class intervals.* Class \$25 to \$30, which is one class interval below (in size) the *assumed mean,* therefore has a deviation of -1. Likewise, the group \$40 to \$45, two class intervals above the origin, has a deviation of $+2$. This process is continued for all intervals. These deviations are shown in column E.

With the determination of the deviations it is necessary to multiply each by the frequency of their respective classes. This procedure results in the product of frequencies and deviations, which may be expressed in symbols, as $(f \times d)$. The mean, however, is the point at which the sum of the deviations equals zero. The deviations must, therefore, be summated by symbols $\Sigma(f \times d)$.

In Table XXXIV column G it is found that the sum of the deviations is not zero, there being an excess of -25, $(-88 + 63)$. The actual mean therefore differs from the *assumed mean,* \$32.

It can readily be seen that the mean has been assumed at too high a figure. If it had been assumed at \$27, the midpoint of class \$25 to \$30, the deviations below the assumed mean would have been smaller than those for the origin at 32. Group \$20 to \$25 would have had -1 deviation instead of -2; group \$15 to \$20, -2 in place of -3. Therefore, by changing the assumed mean to 27 the product of the frequencies and deviations would have been smaller than when the origin was at 32. Conversely, for the same reason, the product of deviations and frequencies above the assumed mean would have been greater. It is evident that by changing the arbitrary origin from 32 to 27 the sum of the product of the minus deviations and frequencies, which is now in excess, would have been smaller; while that for the plus, which is now too small, would have been greater. The *assumed mean* 32 may, therefore, be corrected by working back the arbitrary mean in such a way as to reduce the minus and increase the plus deviations until their sum equals zero.

In a total frequency of 103 with an *assumed mean* of 32 the excess of minus deviations over the plus is -25, $(-88+63)$. The origin, therefore, is really $^{25}/_{103}$ *class intervals too high*. The mean, however, is always measured in the same unit as the variable. The correction, it should be observed, has been measured in class intervals, a unit of \$5. It must therefore be multiplied by the width of the class (C). The correction then becomes $^{25}/_{103} \times 5$ or, symbolically, $\frac{\Sigma(f \times d)c}{N}$.

The entire formula for the interval method may, therefore, be written, as derived from the foregoing paragraphs, as

$$\text{Mean} = A + \frac{\Sigma(f \times d)c}{N}$$

by substitution for the problem

$$\text{Mean} = 31 - {}^{25}/_{103} \times 5 = \$30.79.$$

70. Points to Consider in the Use of the Interval Method.—The interval method, in contrast to the so-called simple method, is relatively simple and a decided time-saving device, and should be used wherever possible. In using it, however, it should always be borne in mind that it is *a class interval method*. all measurements being made in class intervals. Therefore, even though a class interval has no frequency, as class \$40 to \$45 in Table XXXIV, it must always be given a deviation, otherwise the computation fails.

In a series having a miscellaneous group listed as "40 and over," etc., it becomes evident that its midpoint is indeterminate and therefore cannot enter into the computation of the mean.

The arbitrary origin may be assumed in any interval because the excess of minus or plus deviations will indicate whether to subtract from or add to the assumed mean.

This method should never be used when the class intervals of a series vary in width. The method may be used with certain modifications,[1] but the results will be a mean derived from the various means of the unequal groups and therefore not an exact measure of the mean. In such a series it is better to compute the mean either by the so-called simple method, or to arbitrarily assume a mean and measure exact deviations in the actual unit of measurement from the origin and correct to derive the actual mean. Such a method is illustrated in Table XXXV.

TABLE XXXV.—WAGES IN FACTORY Y MEASURED TO NEAREST DOLLAR

Wages A	f B	Deviation from assumed mean \$34.50 C	Product $f \times d$ D	
\$15–\$20	22	−17.50	−385	
20– 25	21	−12.50	−262.5	
25– 30	22	− 7.50	−165	
				−812
30– 40	16	Assumed average		
40– 50	10	+10	+100	
50– 75	12	+27.50	330	
				+430
Totals.....	103	.....		−382

Assumed mean = \$34.50
Correction = $-382/103$ = 3.708
Real mean = \$30.79

In this table \$34.50 is *arbitrarily assumed* to be the mean. From this figure the *actual* deviations of the classes are recorded (column C). Group \$20 to \$25, with a midpoint of \$17, deviates \$−17.50 from the assumed mean; group \$50–\$75, with a midpoint of \$62, has a deviation of \$27.50. The sum of the

[1] See SECRIST, H., "An Introduction to Statistical Methods."

products of these deviations and their respective frequencies yields the correction for the assumed mean. In a frequency of 103 the sum fails to balance by −382 (column D) the correction, therefore, is $-3\frac{82}{103}$, resulting in a mean of $30.79.

This method is a laborious one, but the most suitable in tables with unequal intervals. This statement should emphasize the need for the construction of tables with equal intervals.

71. The Weighted Arithmetic Mean.—All frequency series are weighted to a certain extent, as the mean is a measure which gives weight to the extremes; the amount of the weight depends largely upon the frequency within the group and the distance of the group from the mean. However, in some series such weighting is too haphazard, and in some cases, as in single frequency, there is no differential weighting. Conscious weighting may therefore be necessary.

Weighting is of extreme importance in the making of index numbers, although precision of weighting is not so important as accuracy in the enumeration of the data. In making an index to measure the cost of living, it is self-evident that all items of consumption should not be given the same weight. Condiments have a place in family budgets, but their consumption, hence importance, is negligible when compared with the staple products, as flour, etc. The amount of the budget spent on fuel, clothing, rent, lighting, and food differs, therefore, if cost of living is to be measured by an index the percentage of each to the total budget expenditure would be a good system of weights. Irving Fisher illustrates weighting in the following:

> If sugar is marketed to the extent of ten billion dollars a year, while salt pork is marketed at only five billion dollars a year, there is clearly ground for regarding sugar as twice as important as salt.[1]

Weighting, then, is really assigning special frequencies to the data, the frequencies being assigned according to the relative importance of each datum. The weighted arithmetic mean is therefore computed by dividing the sum of the products of weights and frequencies by the total weights, or by formula

$$Me = \frac{\Sigma(V \times w)}{\Sigma w}.$$

The problems connected with weighting will be given further consideration in Part III.

[1] The Making of Index Numbers," p. 45.

72. Advantages and Limitations of the Arithmetic Mean.—*a. Advantages.*—1. The mean is rigidly defined and is entirely a mathematical concept; therefore it lends itself to algebraic treatment.

2. If the study requires that emphasis be given to extreme items, the mean should be used, as it gives weight to extremes according to frequency and distance from the mean.

3. The mean is based upon all the observations made and is not so apt to be affected by errors in the samples chosen, because by including all observations the errors may be compensating and not confined to all data.

4. The mean is relatively easy of calculation, although not so easy as either, the median or the mode. It may be calculated from the aggregate of items, provided the number of items are given. In other words, the size of each datum is not necessarily required, only the sum of the individual items and their number being necessary; *e.g.*, given the total income of the United States and the total population the mean income may soon be derived.

b. Limitations.—1. A computed mean may be a fictitious value and not actually represented in the series. The mean for Table XXXIV was found to be $30.78, while if it were for weekly wage rates there would be no datum corresponding to the mean, since such wage rates advance by 25-cent gradations and not odd cents.

2. For some studies, because the mean is a mathematical concept, it cannot be used—it cannot measure qualities, as can the median.

3. In some problems the weight given to extremes would be a disadvantage. In studying the distribution of wealth, the relative few who have large returns would pull the average up to a higher limit than if the mean were computed from the more numerous groups. This idea is illustrated in Chart 46, where the addition of an extreme item, a 200-pound weight, makes the average (165) higher than if the average had been computed from the first six persons with an average of 158.

4. The mean cannot be determined graphically.

5. A series cannot be characterized by mere computation of the mean. Two series may be identical in this measure and yet differ considerably in other respects, as illustrated in Chart 47. It should be noted that the polygon designated by the broken line is that for a series with unequal intervals.

In spite of the above limitations the mean is one of the most useful measures of position. Its use, however, should be attended with caution.

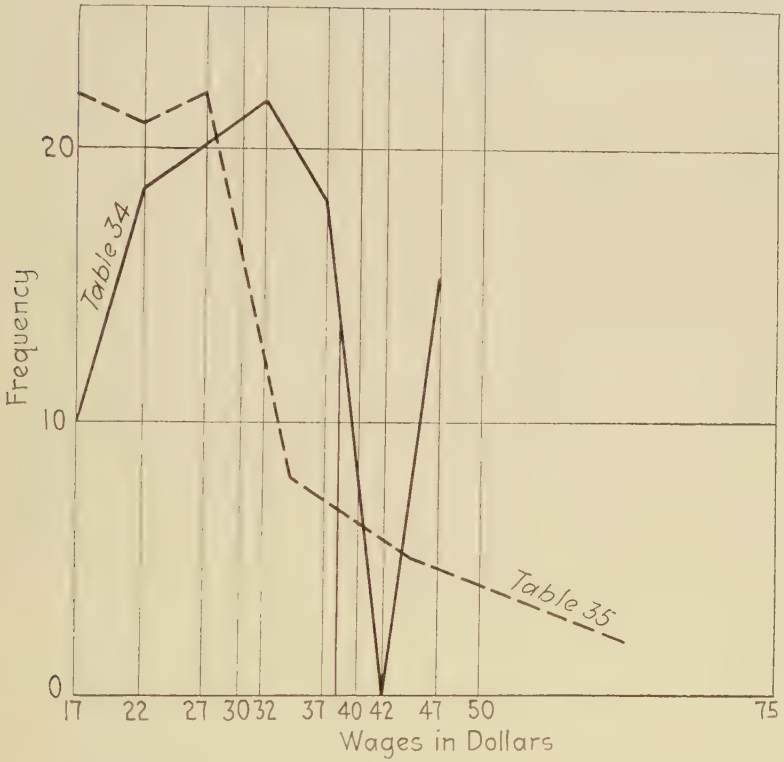


CHART 47.

II. THE GEOMETRIC MEAN

The geometric mean is rarely used in the analysis of ordinary frequency series because of the complexity of its computation in contrast to the arithmetic mean, the geometric involving the use of logarithms. In series, however, where it is necessary to measure ratio changes the geometric mean is used. Its use is confined largely to index numbers, which will be discussed later.

The geometric mean, while based upon all the observations made, does not give as much weight to extreme items as the arithmetic mean, because it measures deviations according to their ratio *to the mean,* whereas the arithmetic measures devia-

tions in ratios according to *their difference from the mean.* From this it is apparent the geometric mean is always smaller than the arithmetic mean.

The method of computation may be summarized by the following formula:

$$\text{Geometric Mean} = \sqrt[n]{V_1 \times V_2 \times V_3 \ . \ . \ . \ V_n},$$

which indicates that the variables are multiplied together and the root of their number extracted from their product.

It may perhaps best be explained by application to the data of Table XXXIV (page 139), in which the logarithms of the respective midpoints are found. These logarithms when multiplied by their respective frequencies, the products being summated, yield the sum from which the root corresponding to the total frequency must be extracted, which results in the geometric mean. The products of logarithms and their frequencies are summated because in logarithms multiplication is achieved by addition. The geometric mean of Table XXXIV becomes

$$\sqrt[103]{(10 \log 17 + 19 \log 22 + 20 \log 27 + 21 \log 32 + 18 \log 37 + 15 \log 47)}.$$

Computation

	Logarithm	Product
10 log 17	1.2304 × 10	12.3040
19 log 22	1.3424 × 19	25.5056
20 log 27	1.4314 × 20	28.6280
21 log 32	1.5051 × 21	31.6071
18 log 37	1.5682 × 18	28.2276
15 log 47	1.6721 × 15	25.0815
Total	.....	151.3538

$$\text{Log of geometric mean} = \frac{151.3538}{103} = 1.4694$$

Antilog of 1.4694 = 29.47 the geometric mean
Geometric mean = $29.47.

The arithmetic mean for the same table was found to be $30.78. The difference from the geometric mean indicates the properties of the geometric mean. Therefore in problems in which too great weight should not be given to extremes but

an analysis based on all the observations made, this mean should be used. Its disadvantage when compared to the arithmetic mean lies in the complexity of computation and therefore is rarely used in frequency tables.

III. THE HARMONIC MEAN

The harmonic mean is another average which receives little attention in the analysis of the general type of frequency series. Its use is found largely in the analysis and construction of index numbers, and even in this field it has a limited use. It is of great value when the unit desired to be measured varies in size, not being constant.

This mean is obtained by computing the arithmetic mean of the reciprocals of the items and then taking the reciprocal of the resultant mean.

If over a series of years an article was quoted as so many pounds to the dollar, that is, the price quotation was constant but the quantity it would purchase varied—in other words, the price per unit varied—it would be difficult from unit figures to compute the mean *price per pound*. The mean computed from the figures would be the mean quantity that could be purchased for a fixed price. As an illustration, suppose an article X was sold in the following quantities per dollar:

	1918	1919	1920	1921	1922	1923
Units per dollar.....	14	18	19	15	14	13

The mean quantity purchased for the years 1918 to 1923 for $1 was 15.5. The problem however, is to find the number of pounds per dollar that could be purchased at the mean price. This necessitates the determination of the mean price per pound and then its reciprocal, or the harmonic mean.

The mean price per pound may be obtained by taking the mean of the reciprocals of the quantities for the successive years, *i.e.*, cost per pound in 1918 = $\frac{1}{14}$ = 0.0714, so that the above figures read:

	1918	1919	1920	1921	1922	1923
Cost on dollars....	0 0714	0.0555	0.0526	0.0666	0.0714	0.0769

The mean of these figures, 0.0657 yields *the mean cost* 0.0656, or cost in dollars and is the mean of the reciprocals of the units per dollar, but by definition the harmonic mean is the reciprocal of this mean. Therefore the

$$\text{Harmonic mean} = \frac{1}{0.0657}, \text{ or } 15.2.$$

This figure, 15.2, indicates that such number of pounds could be purchased for $1 at the mean price per pound.

References

Bowley, A. L.: "Elements of Statistics," pp. 107–118, 128–129.
——: "An Elementary Manual of Statistics," pp. 15–20.
Davies, G. R.: "An Introduction to Economic Statistics," pp. 20–21.
Elderton, W. P. and E. M.: "Primer of Statistics," chap. II.
Kelley, Truman L.: "Statistical Method," pp. 45–53, 63–68.
King, W. I.: "Elements of Statistical Method," pp. 132–140.
Secrist, H.: "An Introduction to Statistical Methods."
——: "Readings and Problems in Statistical Methods," pp. 318–325.
West, Carl: "Introduction to Mathematical Statistics," pp. 32–38.
Zizek, Franz: "Statistical Averages," part II, chap. I–III.

Problems

1. What is the arithmetic mean?

2. Compute, by the interval method, the arithmetic mean for each of the three wage series found in problem 10, Chap. VII.

3. What effect, if any, has the frequency in the class interval $15 to $15.50 of the highest wage series, of the same problem above, upon the value of the mean? Why?

4. From your answer to problem 3, when would you refrain from using the mean to characterize a series?

5. How would you compute the mean for a series which has unequal class intervals?

6. *a.* How would you compute the mean for a weighted series?

b. Why or when would weighting be resorted to?

7. How do the relative values for the arithmetic and geometric mean for the same series compare? Why?

CHAPTER XI

THE MEDIAN, QUARTILES, AND PERCENTILES

I. THE MEDIAN

73. Introduction.—In a frequency series in which the total frequency is an odd number, the median is the measurement of the median item. The median item is that which divides the total number of items, when arranged *in ascending order*, into two equal parts. But all total frequencies are not *odd* in number, so that the middle of the series may fall *between* two central items. *The median, therefore, is that position on the scale, which divides the items, when arranged in ascending order, in such a way that there are as many above as below the central position.* It differs from the mean in that it appears to be primarily interested in the frequency of the variables, whereas the mean divides the deviations of the variables into two equal parts. The median, however, divides the deviations in measurements in such a way that their sum, *if signs are ignored*, is a minimum. This is illustrated by the weights given in Table XXXIII.

149, 154, 154, 160, 163, 165, 200 pounds.
First, second, third, fourth, fifth, sixth, seventh persons.

If the fourth item assumed to be the median item (160 pounds), the deviations are: $-11, -6, -6, 0, +3, +5, +35$, a sum of 66 (if signs are ignored.) If 154 were assumed to be the median, the deviations would be: $-5, 0, 0, +6, +9, +11, +46$, the sum of which, ignoring signs, is 77. Any other value assumed for the median would likewise yield a larger sum than 66, which is derived when the median is the fourth item.

The median position is found by adding 1 to the total[1] frequency and dividing by 2, or by formula $\frac{n+1}{2}$. The necessity of adding 1 may be shown by the above illustration. If the total

[1] For argument against such practice see Chaddock, R. E., "The Graphic Representation of a Frequency Distribution," *Amer. Stat. Pub.*, June, 1921.

frequency of 7 (persons weighed) were divided by 2 the median position would be $3\frac{1}{2}$, that is, between the third and fourth persons, if they were lined up according to the ascending order of weights. Whereas by adding 1 to the total frequency $\frac{7+1}{2}$, the fourth person would be the median item and would divide the frequency into two equal parts, there being three persons on either side of him. The same reasoning would apply if the total frequency were an even number; if eight persons were weighted instead of seven.

1 2 3 4 *M* *m* 5 6 7 8

If the frequency were divided by 2 the median item would be the fourth (*M*), there being four lines to the right and only three to the left. By adding 1 to this frequency $\frac{8+1}{2}$, the median would be represented halfway between 4 and 5 (*m*), the correct midpoint. The rule for obtaining the median item is, therefore, to add 1 to the total frequency and divide by 2; by formula $\frac{n+1}{2}$. The median is then obtained by measuring the median item or position.

74. Computation of the Median.—Data when grouped into homogeneous classes and arranged in ascending order lose much of their detail, so that it may be difficult to find the measurement of the median item. Its measurement may be obtained by interpolation, a measuring within the known median group. For discrete series too great an accuracy should not be expected by such a method because interpolation implies an *equal* distribution, over the interval, of the frequencies within the group. Such distribution is only found in continuous series. However, if discrete series have small intervals, interpolation may be resorted to.

In order to develop the method of computation Table XXXVI is presented. The median item is the seventy-seventh. The problem of finding the median becomes one of locating and measuring this item. By the construction of a "less than" cumulative frequency table, as in column C, the median group may be located. Column C indicates there are 72 items below the class 66 to 68 and 117 items "less than" the upper limit of this class: the seventy-seventh item, therefore, is located within this class. The problem now becomes one of measuring within (interpolation)

TABLE XXXVI.—HEIGHTS OF STUDENTS IN CLASS X
Measured to nearest half inch

Height, in inches A	F B	C.F. C
54–56	3	
56–58	6	9
58–60	9	18
60–62	10	28
62–64	18	46
64–66	26	72
66–68	45	117
68–70	21	138
70–72	10	148
72–74	5	153
	153	

$$\text{Median item} = \frac{153 + 1}{2} = 77$$

$$\text{Median} = 65.75 + \frac{5}{45} \times 2$$

$$= 65.97.$$

the group for the seventy-seventh item. Since measurements are tabulated to the nearest half inch, for tabulating purposes the criterion for the placing of line case data becomes the quarter of an inch. That is, data measuring 65.75 inches would be tabulated in the 66 to 68-inch group, while those measuring 65.70 would be thrown into the 64 to 66 grouping. Therefore the lower limit for class 66 to 68 is 65.75 inches. In the median group there are 45 items which are presumed to distribute themselves regularly over the 2-inch interval since the series is a continuous one. Each successive frequency, therefore, increases over the previous one by $\frac{2}{45}$ inch. The median item must of necessity be the fifth item within the median *group*, since 72 items occur below the class and the median item is 77. Since each item within the class increases over the previous one by $\frac{2}{45}$ inch, the fifth within the class must have increased $5 \times \frac{2}{45}$ ths over the lower limits; hence the median is $65.75 + \frac{2 \times 5}{45}$ inches, or 65.97 inches, which for practical purposes should be called 66 inches.

A formula may be used to indicate the method used in the above illustration

$$\text{Median} = l + \frac{i \times c}{f},$$

in which l = the lower limit of the median group.
i = the number of items to interpolate for within the group (the median item—the total frequency below the median or group).
c = the class interval.
f = the frequency of the median group.

This process of interpolation should be used with extreme care. In a discrete series with large intervals, it is wrong to assume regularity in the distribution of the class frequencies.

Ogive of Data in Table XXXVI

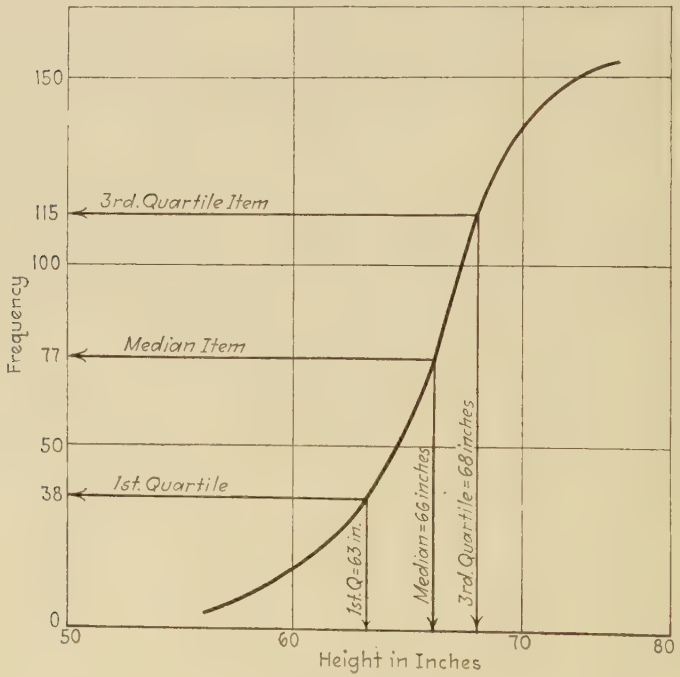


Chart 48.

75. Advantages and Limitations of the Median.—*a. Advantages.*—1. The median is not necessarily a mathematical concept; qualities as well as quantities may be given median ranking.

2. The median is based upon all the observations made and may be computed without knowing anything about the variables at the extremes other than their number and arrangement in ascending order. Measurements of the variables immediately around the median are all those required for computation.

3. The median does not give weight to extreme items. It varies little with the addition or subtraction of an extreme item. In Table XXXIII (page 136) the addition of another person, irrespective of his weight, would change the median only 2 pounds.

4. The median is easier to compute than the mean.

5. The median (quartiles, deciles, and percentiles) may be computed by graphic methods, by means of a cumulative frequency curve or ogive. Chart 48, based on Table XXXVI illustrates the method—how, by finding the median or quartile item on the ordinate, its measurement may be read off the chart at the points of intersection at the curve and abscissa.

b. Limitations.—1. In a series the median may be found only by interpolation, so that if the series is discrete the median so derived may be ficticious.

2. The median, unlike the mean, does not lend itself to algebraic treatment.

3. Some studies require that weight be given to extremes; this the median fails to do.

II. QUARTILES, DECILES, AND PERCENTILES

76. Introduction.—The quartiles, deciles, and percentiles are all measures of position and are used primarily to determine the nature of the frequency curve or distribution. They aid in characterizing the distribution and, as their terms indicate, they are of the same species as the median.

The quartiles divide the total frequency of a series into four equal parts: the first, second, third, and fourth quartiles. The deciles divide the total frequency into ten equal parts called "deciles," while the percentiles divide the distribution into one hundred equal parts. The median, therefore, is identical with the second quartile, the fifth decile, and the fiftieth percentile.

77. Method of Computation.—An actual computation will be made for the quartiles, only, as the method is similar for deciles and percentiles, the only difference being in the division

of the total frequency. For quartiles it is four, deciles ten, and percentiles, one hundred. The numerator is multiplied by the particular quartile, decile, or percentile required.

The first quartile divides the sum of the frequencies from the lowest item to the median into two equal parts, and since the median position is $\frac{n+1}{2}$ the first quartile position must be $\frac{n+1}{2 \times 2}$ or $\frac{n+1}{4}$. In Table XXXVI the total frequency is 153; the first quartile items would be $\frac{154}{4} = 38.5$. As in the method for the median, it is necessary to interpolate for the measurement of this item. Since the frequency 38.5 falls in the group 62 to 64 inches and is 10.5 items, within this 2-inch interval, which has 18 items distributed over its width, the first quartile must equal $61.75 + \frac{10.5 \times 2}{18}$, or 62.91 inches.

The third quartile is the measurement of the item which divides the frequencies above the median into two equal parts; the third quartile item is $\frac{3(n+1)}{4}$, or the item which is three-quarters of the way up the total frequency. In Table XXXVI this item proves to be $\frac{3(154)}{4}$, or 115.5. By the same process as used in the median, the third quartile class is found to be 66 to 68 inches, the item being the 43.5 in a class of 45 which are spread over 2 inches, so that the third quartile equals $65.75 + \frac{(43.5 \times 2)}{45}$, or 67.68 inches.

There is no short cut in the computation of the quartiles. They cannot be derived by dividing the range between the lowest item and the median item by 2, and subtracting this from the median in order to obtain the first quartile or by adding to obtain the third. If such a method were attempted in Table XXXVI, the results would be very different from those already obtained. The range of the lower half of this table is from 53.75 to 65.97 inches (the median), which is 12.22 inches, one-half of which 6.11. If deducted from the median this would give the first quartile a value of 59.86 inches, whereas the computed quartile was found to be 62.75. The reason for this difference is obvious—the quartile figure depends upon the degree of concentration of the frequencies around the norm. If the degree is high, the quartiles will be very close to the median, in

size; if low, they will be more distant. This also explains why the distance between the first quartile and the median is not identical—unless it is a normal distribution—with that between the median and third quartile; hence the quartiles are used as measure of skewness.

The quartile is a good measure of the distribution, as it shows the limits between which the center half of the total frequency of a table may be found. In the illustration, the quartiles indicate that the heights of one-half of the students range from 62.91 to 67.68 inches.

The graphic derivation of the quartiles is indicated in Chart 48.

With an understanding of the method in computing the median and quartiles there should be no difficulty in the derivation of deciles and percentiles.

References

Bowley, A. L.: "An Elementary Manual of Statistics," pp. 23–24.
——: "Elements of Statistics," pp. 124–128.
Davies, G. R.: "An Introduction to Economic Statistics," pp. 27–32.
Elderton, W. P. and E. M.: "Primer of Statistics," pp. 1–22.
Kelley, Truman L.: "Statistical Method," pp. 54–59.
King, W. I.: "Elements of Statistical Method," pp. 127–132.
Secrist, H.: "An Introduction to Statistical Methods."
West, Carl: "Introduction to Mathematical Statistics," pp. 38–42.
Yule, G. U.: "An Introduction to the Theory of Statistics," pp. 116–120.
Zizek, Franz: "Statistical Averages," part II, chap. IV.

Problems

1. What are the characteristics of the median?

2. Compute the medians for the three series of problem 10, Chap. VII.

3. Does the class $15.50 of the highest wage group affect the value of the median as it did the mean? Why?

4. Are the differences between the respective medians smaller or larger than the differences between the respective means of the above series? Why?

5. For the same series, compute the first and third quartiles.

6. Compare the differences, for the respective series, of the respective quartiles from their medians.

7. *a.* Why may the quartiles for any one series differ in their distance from their median?

b. How could you explain the variations between the series in distance of the quartiles from the respective medians?

8. From the table used above, compute the second and eighth decile for the lowest wage group.

9. Derive graphically from the ogives drawn in problem 5, Chap. VIII; the median, first, and third quartiles, and the second and eight deciles for the highest wage group.

CHAPTER XII

THE MODE

78. Introduction.—The mode is the measurement of the item of greatest frequency. The mode for Table XXXIII is 154, as there are two persons of this weight in contrast to single frequencies for the other weights. The mode may be read from a frequency curve, and is the measurement of the item represented by the peak of the curve, and is the item of greatest density. The mode is perhaps the average commonly used, usually termed an average; the average size shoe usually means the size which has the greatest demand and is really the mode.

The modal group may be determined by inspection, the group which has the greatest frequency, but if there is no definite point of concentration the method known as moving averages must be applied. The measurement of the modal item within the group is determined usually by interpolation, although it may be derived by the application of the Pearsonian formula. These methods will now be discussed.

79. Moving Average.—The moving average has to be resorted to when series, through defective arrangement of intervals (they being too small), fail to show any one point of concentration.

In Table XXXVII there is no definite point of concentration, the series being decidely bimodal. In column 1 there is no definite massing of the frequencies, the maximum frequency 17 being found in two classes which are 4 centimeters apart. Seemingly, there are two modes for this series. One definite modal group may be derived by means of the moving average. In this method the class intervals are constantly widened and the limits changed, in order to locate the class which will maintain itself within the larger frequency groups.

By enlarging the interval to 3 instead of 1 centimeter the material begins to mass, with a maximum of 40 for the group 160.5 to 163.5, column 2. This group of largest frequency, however, is made up of three class intervals. Therefore, the modal class cannot be determined until the limits and group-

TABLE XXXVII

Height in centimeters	Frequency	Frequency groupings									
		1	2	3	4	5	6	7	8	9	10
154.5–155.5	4										
155.5–156.5	13	25									
156.5–157.5	8		30		34						
157.5–158.5	9			28		41			45		
158.5–159.5	11	33					41			54	
159.5–160.5	13		34					43			51
160.5–161.5	10			40	51						
161.5–162.5	17	40				53			58		
162.5–163.5	13		35				45			59	
163.5–164.5	5			32				49			66
164.5–165.5	14	36			49						
165.5–166.5	17		43			48					
166.5–167.5	12			36			50		55		
167.6–168.5	7	24						41			
168.5–169.5	5										

ing of the intervals have changed so as to permit the location of the class which maintains itself in the largest frequency groupings. By moving the limits up, in size, one class interval, column 3, the modal group has a frequency of 43 and the grouping excludes classes found in the maximum group; therefore, the change in grouping must be continued. Column 4 brings the modal group back to such limits as include two classes (160.5. to 162.5) which were found in the first large grouping column 2. Column 5, indicates groupings of 4 centimeters, the maximum frequency of 51, which still includes the intervals 160.5 to 162.5 In column 8, 161.5 to 162.5 is clearly the only original class to maintain itself within the largest group. By continuing the process this interval is found to be the modal group. The mode may now be determined by interpolation.

This method is tedious, but the only one possible to use in series having bimodal tendencies.

80. Interpolation.—In finding the median it was necessary to interpolate for the median item. This method is also used in the determination of the mode; within the modal group. *Interpolation should be used only when the series are continuous*, the

mode should always be actually represented in the series and not a fictitious value.

In Table XXXVI the modal group is, as is self-evident, 66 to 68 inches, having a frequency of 45. Within this must be measured the modal item. In most series the frequencies concentrate around the mode. The frequencies above the modal group tend to pull the mode above the lower limit of the class 66 to 68, while those below tend to pull it below the upper limit. This suggests a method of interpolating for the mode, which may be found by finding the relationship between these two tendencies or pulls. It may derived by dividing the sum of all the frequencies *above the modal group* by the sum of the frequencies *above and below, not including* the modal group. However, for ordinary calculation, in that the frequencies tend to mass around the mode, the frequencies of the class immediately above and below the modal group will suffice. Applying this to Table XXXVI, there are 21 items which tend to pull the mode above 65.75 inches, the lower limit of the class 66 to 68, and 26 pulling it down from the upper limit, in an interval of 2 inches. The mode therefore becomes

$$65.75 + \frac{21 \times 2}{26 + 21} = 66.64 \text{ inches.}$$

By formula the mode may be represented

$$\text{Mode} = l + \frac{fc+ \times c}{fc- + fc+},$$

in which l = lower limit of the modal group
c = class interval.
$fc+$ = frequency of class above modal group.
$fc-$ = frequency of class below modal group.

This is a relatively simple but accurate method of finding the mode if applied to continuous series. For discrete series it should be used with caution.

81. Pearsonian Formula.—The mode may be computed for series which have only moderate skewness by means of its relationship to the mean and the median. Pearson devised the following formula, which serves as a rough measure of the mode:

$$\text{Mode} = \text{Mean} - 3(\text{Mean} - \text{Median}).$$

82. Advantages and Limitations of the Mode.—*a. Advantages.*

1. It is the most typical item of a series and is the average usually used by the layman.

2. It does not emphasize extremes and, unlike the median and mean, is not based upon all the observations made. It is calculated from the frequencies immediately above and below the modal group.

3. It may be located graphically.

b. Limitations.—1. A series may be bimodal, showing no definite point of concentration. The mean and the median, on the other hand, are definite.

2. If the extremes should be emphasized the mode cannot be employed.

3. All its computations are subject to great inaccuracies when resort is made to interpolation.

4. It gives little information with regard to the general character of a series, its computation being confined to but three classes in a table.

5. In computation emphasis is placed upon the frequencies and not the variables themselves, with the exception of the mode itself.

References

Bowley, A. L.: "An Elementary Manual of Statistics," pp. 20–23.
——: "Elements of Statistics," pp. 118–124.
Davies, G. R.: "An Introduction to Economic Statistics," pp. 21–27.
Elderton, W. P., and E. M.: "Primer of Statistics," pp. 40–46.
Kelley, Truman L.: "Statistical Method," pp. 60–62.
King, W. I.: "Elements of Statistical Method," pp. 122–127.
Secrist, H.: "An Introduction to Statistical Methods."
West, Carl: "Introduction to Mathematical Statistics," pp. 142–44.
Yule, G. U.: "An Introduction to the Theory of Statistics," pp. 120–123.
Zizek, Franz: "Statistical Averages," part II, chap. V.

Problems

1. What are the characteristics of the mode?

The following problems are based upon problem 7, Chap. X.

2. Does the frequency of class \$15 to \$15.50 of the highest wage group affect the computation or value of the mode? Why?

3. What procedure should be adopted in the determination of the mode for the highest wage series? Why?

4. Determine the mode, by the interpolation method, for each of the three series.

5. Compare the values of the modes derived above to those for the same series derived by the Pearsonian formula.

6. Derive the values of the mode graphically from the polygons drawn in problem I, Chap. VIII.

CHAPTER XIII

SUMMARY ON AVERAGES

In the preceding four chapters consideration has been given to the various measures of position, together with the advantages and limitations of each. The latter are by no means fixed and what might be considered an advantage for one problem may prove a disadvantage for another. There is no hard-and-fast rule determining the average to use. The purpose of the study, however, to a considerable extent, determines the measure of position to use. The storekeeper in selling shoes stocks his store largely by studying the mode. The factory manager, in reviewing the efficiency of the plant for the past year, measures continuous production and not merely the slack and busy seasons. He therefore computes the mean for the year, which is based upon all days of production, and gives weight to the marked changes in output. The use of the mean in an income study, by giving weight to extreme items, may vitiate the object of the study, whereas the quartile range would show the limits within which one-half of the income receivers may be found.

It is impossible wholly to characterize any series by computing but one of the averages. The problem may only demand one measure, but in such cases the series should not be considered to be alike in other respects. Likewise, when two series have one identical measure it should not be assumed that the series are identical in other respects. This is evident from the analysis of Tables XXXIV and XXXV, which are summarized below.

TABLE XXXVIII

Table	Mean	Median	Mode	Quartiles	
				First	Third
XXXIV	$30.79	$30.21	$31.86	$23.71	$36.72
XXXV	30.79	26.55	Bimodal	20.45	37.62

Comparison shows, in Table XXXVIII, these two series to be radically different from one another in all respects save the means, which are identical. Table XXXV has no definite concentration in frequencies and therefore is bimodal. The differences between the respective medians and quartiles of the series are considerable. The quartile measures indicate the curves to have totally different conformations, as illustrated in Chart 47 (page 145). This emphasizes the fact that averages used to characterize a series should be used with extreme care, since values so derived may be fictitious. Averages are at all times only one of the many possible pictures of the series.

Averages in themselves are usually *insufficient*. The storekeeper has to have some shoes that are larger or smaller than the modal size. How many should he have? The factory manager has to provide for seasonal activity. How large a plant is necessary? These questions cannot be answered by any one average or by all the averages combined. There are other elements in a series which should be brought out. The scatteration or dispersion and the points of concentration or skewness become important elements in some problems and will be discussed shortly. They are mentioned here in order to illustrate the narrow limits within which averages should be used. Averages are good tools, and become effective when used under the right conditions and combined with other implements.

TABLE XXXIX.—SUMMARY OF TABLES USED IN CHAPS. IX, X, XI, AND XII

Table	Page	Mean	Median	Mode	Quartiles	
					First	Third
XXXII	134	$ 0.70	$ 0.70	$ 0.68	$ 0.65	$ 0.75
XXXII	134	0.60	0.58	0.58	0.56	0.63
XXXIV	139	30.79	30.21	31.86	23.71	36.72
XXXV	142	30.79	26.55	Bimodal	20.45	37.62
XXXIII	136	165 pounds	160 pounds	154 pounds		
XXXVI	151	65.18	65.97	66.64	62.91	67.68
XXXVII	157	162.2	161.6	162.1	159.159	166.6

Problems

1. What are the requisites of a good average?

2. Classify the various averages according to the requisites as derived from problem 1.

3. When, and to what degree, should interpolation be resorted to?

4. Under what circumstances would you use each of the respective averages?

5. Averages are measures of position. Explain.

6. To what degree of refinement should the computed averages be given?

CHAPTER XIV

DISPERSION

83. Introduction.—As stated in the last chapter, the measure of "central tendency" is not always the important measure of a series. In many social and economic studies the important problem is not that of averages, but how great are the deviations from those averages—the scatteration or dispersion. In a social study of undernourishment the very title suggests that the deviations from the standard constitute the important problem. In a standard of living investigation it is, again, the deviations which may be required, namely, the determination of the number of workers who receive wages lower than will suffice for a decent standard. In the study of incomes the averages may be fictitious values, and the problem at once suggests the question of the distribution of all incomes. All these studies suggest the study of dispersion, which is a measure of the degree of concentration of series around some standard.

The necessity of having some other way to characterize a series besides the measures of position may be apparent from the following tables and chart. The tables show how two series having the same mean, median, mode, and quartiles may differ in dispersion.

In Tables XL and XLI the mean, median, mode, and quartiles are identical, and yet Chart 49 indicates that the two series differ in the degree of concentration. Table XL spreads itself over a range of 32 inches, while Table XLI has a range of only 22 inches. These differences may be important. It becomes necessary, therefore, to devise some means of computing these differences so that the comparison of series may be carried on in all its phases. In these particular tables there is only the problem of dispersion, because, as is evident from the graph, the two series have *a normal frequency distribution* and are symmetrical and have no skewness. Such symmetry is seldom found in the comparison of series, but these tables were devised in order to show the importance of characterizing series in terms other than

TABLE XL

Length in inches	f
0– 1.9	1
2– 3.9	2
4– 5.9	5
6– 7.9	7
8– 9.9	10
10–11.9	12
12–13.9	16
14–15.9	19
16–17.9	20
18–19.9	19
20–21.9	16
22–23.9	12
24–25.9	10
26–27.9	7
28–29.9	5
30–31.9	2
32–33.9	1
	164

Mean = 17 inches
Mode = 17 inches
Median = 17 inches
First quartile = 12.5 inches
Third quartile = 21.5 inches

TABLE XLI

Length in inches	f
6– 7.9	11
8– 9.9	12
10–11.9	14
12–13.9	16
14–15.9	19
16–17.9	20
18–19.9	19
20–21.9	16
22–23.9	14
24–25.9	12
26–27.9	11
	164

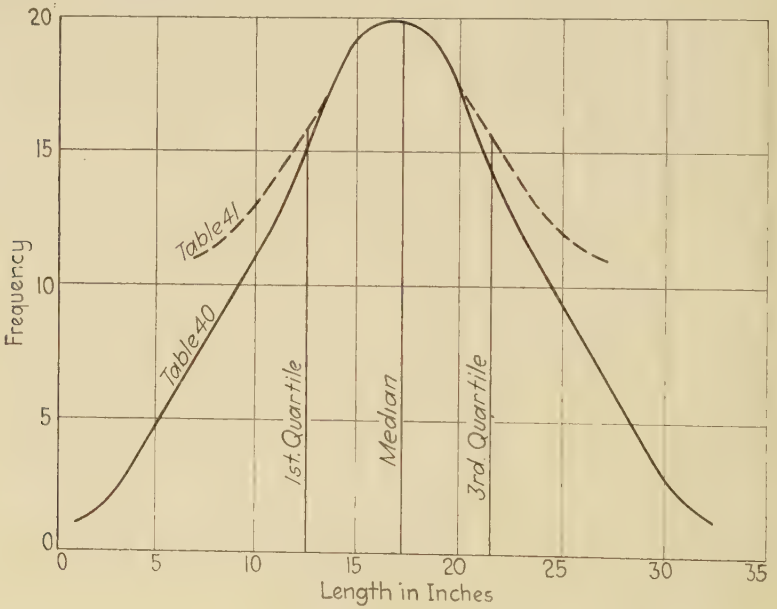


CHART 49.

those of averages; even though all averages are identical for the two series.

84. The Range.—One method of comparing series for dispersion has already been suggested, since the range may be used. This is the spread of the measurements from the smallest item in the series to the largest. However, such a method is a very poor one and only a rough measure of dispersion, since the addition of an extreme item to one of the series would increase its range considerably and yet alter the table very little in other respects, *i.e.*, in Table XL the finding of an actual datum, a freak perhaps, measuring 42 inches would increase the range 10 inches. This is an important defect in the use of the range as a measure of dispersion, but a more important one arises because the range does not show the degree of concentration, that is, the distribution of the frequencies within the range. Two series may have the same range but the frequencies contained within the range may be distributed in a very different manner. Some series may have the frequencies massed around the norm, while others may be bimodal, with no definite point of concentration.

The range, therefore, should be used only as a very rough measure of dispersion, and preferably not at all.

In measuring scatteration it is better to find a relationship to some norm, or "central tendency," because this is usually the point from which the inquiry starts, as seen in the opening paragraph of this chapter. Any one of the averages may be used as the standard from which to measure deviations or dispersion.

85. Measures of Dispersion.—With the selection of the type of average to use as the standard from which to measure dispersion, the next step is the development of the methods of computing it. There are three measures of dispersion:

1. The mean deviation.
2. The standard deviation.
3. The quartile range.

86. The Mean Deviation.—The mean deviation is sometimes called the average deviation. This term is somethat misleading, as an average may be the mean, median, or mode. Although the deviations may be made from any one of the averages, nevertheless the result or measure of dispersion is obtained by taking *the mean* of the sum (*ignoring signs*) of the deviations; hence it is the mean deviation. *The mean deviation, therefore, is*

the mean of the sum of deviations, summated irrespective of signs, the deviations being measured from some standard.

Such a measure is not strictly a mathematical concept, since it violates the algebraic rule with regard to signs. If the standard adopted were the mean, the sum of the deviations would, according to algebra, be zero. Direction of scatteration, however, is not important, that is, whether or not the deviation is below the mean, being minus, or above the mean, being plus. The important thing is the amount of deviation of the items from the standard.

a. The Standard to Use in Computation.—Since the median is the measurement of the item which divides the total frequency of a table into two equal parts, and is a measure which does not give weight to extreme items, it appears to be the standard to use in obtaining the deviations. Furthermore, the median is such that the sum of the deviations, *if signs are ignored*, is at a minimum. This is illustrated below. The sum of the deviations

when the standard is at zero, row 1, is 30. If the standard should be moved to the left one position, row 2, the summation of the left side would be 10 instead of 15; while the right side on the other hand, would have increased from 15 to 21, making a total of 31 in contrast to 30 for row 1. The mean deviation, therefore, if calculated from the median, would be at its minimum.

Notwithstanding the advantages of the median as listed above, the mean is usually used as the standard in preference to the median. As already indicated, the mean has to be resorted to no matter what standard is used, and since the mean lends itself to algebraic treatment it really is the better standard for measuring deviations. By using it homogeneity of treatment is preserved, the mean being used as the standard from which deviations are made, and also the measure used in the derivation of the mean of the sum of the deviations.

b. Actual Deviation Method.—In a single frequency series the computation of the mean deviation is simple—all that is required

is to follow the definition. Below are listed two series which have the same mean but differ in their measure of dispersion.

SERIES A

1, 3, 5, 7, 8, 9, 12, 13, 14. Total of 72 for 9 items. Mean equals 8.
Deviations from mean: $-7, -5, -3, -1, 0, 1, 4, 5, 6$. Total, ignoring signs, equals 32. Row 2.

$$Me.\ dev. = {}^{32}\!/_{9} = 3.55.$$

SERIES B

2, 3, 4, 5, 6, 7, 8, 8, 9, 10, 11, 15, 16. Total of 104 for 13 items. Mean equals 8.
Deviations from mean: $-6, -5, -4, -3, -2, -1, 0, 0, 1, 2, 3, 7, 8$. Total, ignoring signs, equals 42. Row 2.

$$Me.\ dev. = {}^{42}\!/_{13} = 3.23.$$

The first step in the computation of the average deviation is the computation of the mean, the standard from which the deviations are to be measured. Deviations are then calculated from this standard, as shown in row 2 of series A and B. The mean for both series is 8, while the deviations from the standard for the A series comprising 9 items, when summated irrespective of sign, totals 32. The mean deviation, therefore, is 3.55. Series B, by the same procedure, yields a dispersion of 3.23. The comparison of these two series by means of dispersion indicates that series B, although dealing with more items, nevertheless has greater concentration of its data around the standard than series A.

From the illustration it becomes evident that a small measure of dispersion indicates a clustering of the data about the mean and that there are few extreme items. A large dispersion indicates either the predominance of extreme items or little concentration of data about the mean.

The method of computation may be summarized by the formula

$$\text{Mean deviation} = \frac{\Sigma[d-(-d)]}{N},$$

in which d = deviations from the mean.
N = total frequency.

In order to permit the mathematical summation of the deviations, *ignoring signs*, the minus deviations $-d$ are placed within a minus bracket, which, by simplifying, becomes $+d$.

For a frequency table the above formula may be used, the only difference being that the midpoint of the class is taken as

typical of the class and the deviations for these points are calculated from the standard. These deviations, when multiplied by their respective frequencies and their products summated, form the numerator of the formula which becomes then

$$\text{Mean deviation} = \frac{\Sigma fd - (-df)}{N}.$$

If the deviations are measured from the *actual mean* the sum of the deviations from the mean equals zero, the plus being equal to the minus. Therefore, the summation of the deviations, if *the signs are ignored*, must equal twice the sum of the plus deviations. The above formula may now be written as

$$\text{Mean deviation} = \frac{2\Sigma(+df)}{N},$$

in which the deviations and the frequencies *above* the mean are those used in computation.

Table XLII, columns A and B, illustrates this method. The actual mean for this series was found to be 30.79 (page 139). The deviations are shown in column A. The plus deviations, when multiplied by their respective frequencies and the products summated, result in a total of 380.34, as shown in column D. This, when multiplied by 2 and divided by the total frequency of 103, results in the mean deviation of $7.38.

$$\text{Mean deviation} = \frac{2 + 380.34}{103} = \$7.38.$$

c. The Interval Method.—In order to avoid multiplication of large numbers an interval method may be derived. This method may only be used when the *assumed mean* is actually within the real mean group. Therefore, to avoid modifications of the following method it is well, if necessary, to begin the computation by determining the actual mean group. This is not an extra step, as the correction for the assumed mean must be known.

As the title implies, all measurements are made in intervals. Therefore, the result will be in class intervals unless multiplied by the width of the class (C).

In Table XLII the assumed mean is $32 and from this the deviations in class intervals are computed, as shown in column C. Since the mean deviation is the mean of the sum of the deviations (signs ignored) measured from some standard, it requires the respective deviations to be multiplied by their respective frequencies ($f \times d$), column D, and the summation of

TABLE XLII.—TABLE ILLUSTRATING THE TWO METHODS OF COMPUTING THE MEAN DEVIATION

Given mean = \$30.79
Correction for assumed mean = −0.24

Wages	f	Actual deviation method		Interval method	
		A d	B $d \times f$	C d	D $d \times f$
\$15–\$20	10	13.79	.....	3	30
20– 25	19	− 8.79	.....	2	38
25– 30	20	− 3.79	.....	1	20
30– 35	21	1.21	25.41	0	
35– 40	18	6.21	111.78	1	18
40– 45	0	11.21	.....	2	
45– 50	15	16.21	243.15	3	45
Total.....	103	.....	380.34	..	151

Actual Deviation Method.

$$= \frac{2 \times \Sigma f \times d}{N}$$

$$\text{by substitution} = \frac{2 \times 380.34}{103} = \$7.38$$

Interval Method.

$$\text{mean deviation} = \frac{[\Sigma(f \times d) \pm K(N_1 - N_2)]}{N} \times C$$

$$\text{by substitution} = \frac{151 - 0.24(49 - 54)}{103} \times 5$$

$$= \frac{151 + 1.2}{103} \times 5$$

$$= 1.479 \times 5 = \$7.38$$

these products, $\Sigma(fd)$, to be divided by the total frequency N. In Table XLII such procedure yields $^{151}/_{103}$. The total deviations 151, however, were obtained from an *assumed mean* and therefore do not actually represent the deviations from the actual mean. It may be corrected, however, by applying the correction for the assumed mean. This correction K was found to be $-\frac{25}{0.103}$ or −0.24 intervals, as given on page 140. Since it is a minus correction it indicates that the assumed mean was chosen at

too high a figure, which implies that the deviations (N_1) below the assumed mean group in size were all to high. That is, the sum (49) of the frequencies of the classes below the assumed mean group have been multiplied by deviations which were too large by 0.24. Conversely, all items within the assumed mean group and above it (N_2) 54 in number, have been multiplied by deviations which were too small to the extent of 0.24. From these figures the correction for the sum of the deviations from the assumed mean may be obtained, that is, 0.24×49 must be subtracted, while 0.24×54 must be added to sum of the deviations.

The mean deviation of Table XLII

$$= \frac{151 + 0.24(54) - 0.24(49)}{103} \times 5$$

simplifying

$$= \frac{151 + 0.24(54 - 49)}{103} \times 5$$

by formula

$$= \frac{\Sigma fd + K\,(N_2 - N_1)}{N} \times C.$$

Such a formula may be confusing to the student, as the correction K is always considered plus, even though it is actually minus, as in the above problem. In order to make K according to sign and so avoid confusion the formula may be written

$$\textit{Mean deviation} = \frac{\Sigma fd \pm K\,(N_1 - N_2)}{N} \times C,$$

in which N_2 is always considered minus and N_1 plus. By substitution for Table XLII,

$$\textit{Mean deviation} = \frac{151 - 0.24(49 - 54)}{103} \times 5 = \$7.38,$$

which gives the same result as the other formula.

In the above problem the correction is minus, but if it were plus the same chain of reasoning would be followed to determine the values for N_1 and N_2. It would be necessary to determine the frequencies whose deviations have been assumed too small and those too large. Assuming for the moment that K in table XLII were plus, N_1 would be represented by 70. Since the deviation is being considered plus it would indicate the mean to be *assumed too* low. Therefore, all the frequencies below in size and *including* the assumed mean group would have too small a deviation by 0.24. The assumed mean group was arbitrarily

given a zero deviation, whereas it should actually have a -0.24 deviation. Therefore the frequency in this group is included in the total for N_1. Conversely, all the frequencies above, in size the assumed mean group are too large. Therefore N_2 would consist of the remaining frequencies (33). This reasoning leads to the following rules:

1. When K is plus, N_1 is the sum of the frequencies below (in size) and *including* the assumed mean group.

2. When K is minus, N_1 is the summation of the frequencies below but *not including* the assumed mean group.

The formula in its final form becomes

$$\textit{Mean deviation} = \frac{\Sigma fd \pm K(N_1 - N_2)}{N} \times C,$$

in which Σfd = sum of frequencies times deviations in intervals from assumed mean.

K = correction in intervals for arbitrary mean.

N_1 = summation of frequencies below and including the assumed mean group when K is plus, and the summation of those below the assumed mean group when K is minus.

N_2 = the difference between the total frequency and N_1.

C = the class interval.

CHAPTER XV

DISPERSION (*Continued*)

STANDARD AND QUARTILE DEVIATIONS

I. THE STANDARD DEVIATION

87. Definition.—The standard deviation does not violate any rule of mathematics as found in the computation of the mean deviation where plus and minus signs were ignored in the addition. The standard deviation overcomes this illogical process by squaring all the deviations (d^2). This makes all the signs plus, as minus times minus always gives plus. By so doing the deviations may be added in strict conformity to arithmetic principles. This may be illustrated by the simple example used in the derivation of the mean deviation, series A, which has a mean of 8.

Series A..... 1, 3, 5, 7, 8, 9, 12, 13, 14. Mean = 8.
Deviations f r o m
mean..... −7 −5 −3 −1 1, 4, 5, 6. Mathematic summation = 0.
Deviations squared 49, 25, 9, 1. 1, 16, 25, 36. Total = 162.

Naturally, in squaring the deviations their sum has been enhanced by the square of the various deviations, and when this sum is divided by the number of items ($^{162}{}_{9}$), in order to obtain the mean deviation, the result is too large by the amount of the square of the deviations. Therefore, to reduce this result to the normal figure the square root must be extracted, which equals $\sqrt{^{162}{}_{9}} = \sqrt{18} = 3.7$. This is the standard deviation of the series.

From the above example it is seen *that the standard deviation is the square root of the arithmetic mean of the sum of the squares of the deviations when measured from the arithmetic mean.* Or by formula $= \sqrt{\frac{\Sigma d^2}{N}}$.

88. Computation.—Actual Deviation Method.—From the foregoing illustration the method of computation for a single frequency series is apparent. However, most series have varying

frequencies and also class intervals. The problem of measuring dispersion in such series may be resolved into its simplest form by an expansion of the formula $\sqrt{\frac{\Sigma d^2}{N}}$ to include the frequencies. This method is illustrated and the formula derived from Table XLIII.

TABLE XLIII.—TABLE SHOWING THE METHODS OF COMPUTING THE STANDARD DEVIATION. THE ACTUAL DEVIATION AND INTERVAL METHODS

Mean = \$30.79
Given correction for mean = −0.24 interval

Wages	f	Actual deviation method			Interval method		
		C d	D d^2	E $d^2 \times f$	F d	G d^2	H $d^2 \times f$
\$15–\$20	10	13.79	190.16	1,901.64	3	9	90
20– 25	19	− 8.79	77.26	1,467.94	2	4	76
25– 30	20	− 3.79	14.36	287.24	1	1	20
30– 35	21	1.21	1.46	30.74	0	0	0
35– 40	18	6.21	38.56	694.16	1	1	18
40– 45	0	.....	.....	.....	2		
45– 50	15	16.21	262.76	3,941.4	3	9	135
Totals.....	103	.....	.....	8,323.12	..	..	339

Actual Deviation Method.

$$\sigma = \sqrt{\frac{\Sigma f \times d^2}{N}}$$

by substitution $= \sqrt{\frac{8{,}323.12}{103}} = \sqrt{80.80}$ = \$8.99

Interval Method.

$$\sigma = \sqrt{\frac{\Sigma f \times d^2}{N} - K^2 \times C}$$

by substitution $= \sqrt{3\,3\,9\!/\!1\,0\,3 - (-0.24)^2 \times 5.}$
$= \sqrt{3.291 - 0.057} \times 5.$
$= \sqrt{3.234} = 1.798$ intervals.
$\sigma = 1.798 \times 5$ = \$8.99

In this table the actual mean was found to be \$30.79. The first step in the computation becomes apparent, the calculation

of the deviations of the midpoints of the classes from the *actual* mean. These deviations (d) are shown in column C. The deviations are then squared, column D, but the frequencies themselves are important items in determining the degree of dispersion; the deviations squared (d^2) must therefore be multiplied by their respective frequencies ($f \times d^2$), as shown in column E. The total of this column, $\Sigma(f \times d^2)$, measures the sum of the product of deviations squared and frequencies. The standard deviation, however, was found to be the square root of the mean of such a sum. The standard deviation by this method may therefore be derived by the formula:

$$\sigma = \sqrt{\frac{\Sigma(f \times d^2)}{N}}.$$

By substituting values for Table XLIII,

$$\sigma = \sqrt{\frac{8{,}323.12}{193}} = \$8.99.$$

This method, as is evident, is a laborious one and may be simplified by the interval method.

89. The Interval Method.—As the title indicates, all measurements are made in class intervals, and therefore the result must be reduced finally to simple units by its multiplication by the class interval.

In Table XLIII, it should be observed that the arbitrary mean has been assumed at the midpoint of class \$30 to \$35. Deviations, *in class intervals*, are measured from this point, as illustrated in column F; these are then squared, column G, and when multiplied by their respective frequencies yield column H, ($f \times d^2$). The sum of this column gives the sum of the products of the deviations squared and the frequencies. This total, however, can only be correct if the assumed mean happens to be the actual mean. In this table it is found that the assumed mean is 0.24 interval higher than the actual mean; therefore, all the deviations below the assumed mean are really too large, while those above it are too low. Since all the deviations have been squared, all signs have been lost and the only possible correction is to deduct from the sum of the products of frequencies and deviations squared, $\Sigma(fd^2)$, the correction squared (k^2) multiplied by the total frequencies, or $N(K^2)$, since the deviation of every item has erred by K. The correction for $\Sigma(f \times d^2)$ therefore becomes $-N(K^2)$. Since the standard deviation is the square

root of the mean of the sum of the deviations squared times frequencies, the formula for the interval method therefore is

$$\sigma = \sqrt{\frac{\Sigma(f \times d^2) - N(K^2)}{N}}.$$

These measurements, however, are in class intervals, so that the formula must be multiplied by C, the class interval.

In Table XLIII, $\Sigma(f \times d^2)$ equals 339, as shown in column H. $-N(K^2)$ equals $-103(0.24)^2$ while N equals 103. The class interval C equals 5. By substitution in the above formula

$$\sigma = \sqrt{\frac{339 - 103(0.057)}{103}} \times 5 = \underline{\underline{\$8.99.}}$$

In order to avoid unnecessary multiplication (the product of K^2 and N which is deducted from the sum $\Sigma(f \times d^2)$, the difference being divided by N), the formula may be written in its final form

$$\sigma = \sqrt{\frac{\Sigma(f \times d^2)}{N} - K^2} \times C.$$

By substitution for Table XLIII

$$\sigma = \sqrt{339/103 - 0.057} \times 5 = \underline{\underline{\$8.99.}}$$

This method is relatively simple and easy of calculation. One word of caution should be given, however, in the use of the formula—K^2, as the formula indicates, is *always* deducted from the *quotient of* $\Sigma\left(\frac{f \times d^2}{N}\right)$ and the square root extracted from the result before multiplying by C, the class interval.

II. PROPERTIES AND COMPARISON OF MEAN AND STANDARD DEVIATIONS

The mean deviation is not, strictly speaking, a mathematical concept, for in its computation it violates arithmetic rules. Therefore such a measure does not lend itself to algebraic treatment. The standard deviation, on the other hand, by squaring the deviations conforms to all mathematical precepts and permits further algebraic treatment.

The standard deviation, however, in order to conform to algebraic rules, gives additional weight to extremes. This is evident in the comparison of Tables XLII (page 169) and XLIII. In the computation of the mean deviation, Table XLII, the class \$15 to \$20 has a deviation of 3, while the same class in Table

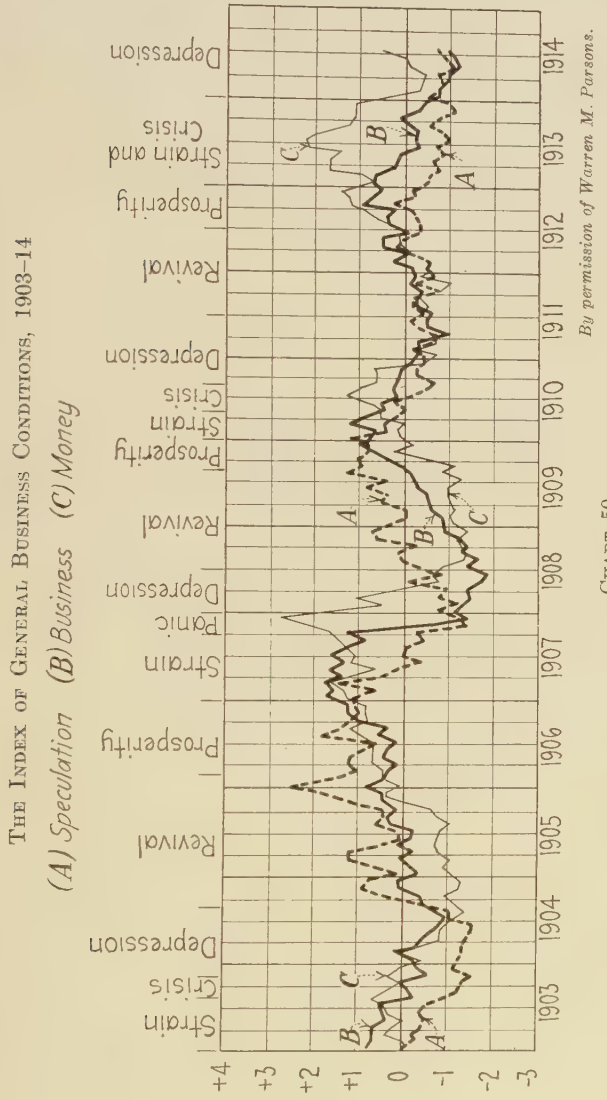
XLIII in the computation of the standard deviation has a deviation of 9; that is, the more distant the item is from the mean the greater is the weight given to it by the process of squaring deviations. The class nearest to the mean group adds nothing to its product by the squaring of the deviation, as 1 squared remains 1. This explains why, for the same series, the measures of dispersion differ—the mean deviation being always smaller than the standard deviation. The mean deviation for Table XLII is \$7.38, while the standard deviation is found to be \$8.99. For moderately skewed series there is a definite relationship between the measure of the mean and the standard deviation. The mean deviation is usually about 80 per cent of the standard deviation. In the tables already referred to this relationship is $\frac{\$7.38}{8.99} = 82$ per cent. Because of the additional weight given to extreme items the standard deviation was a device generally used in the natural sciences but shunned in economic studies. In correlation studies, however, studies endeavoring to show degrees of association between series, the use of the standard deviation becomes imperative, and therefore, together with the arithmetic mean, is now considered by economists as one of the most important measures by which to characterize a series. As with the mean, however, if the problem or analysis does not lead to correlation and does not wish to emphasize extreme items, the average deviation should be used.

The standard deviation is such that, in moderately skewed series, having a fair number of observations, when multiplied by 6 the resulting figure includes "99 per cent or more of all the observations."[1] This may be used as a rough check on the computed standard deviation. It merely means that within a range of six times this deviation 99 per cent of the total frequency of the table will occur. If applied to the tables used in the chapters on dispersion it will be found that such a measure covers more than 99 per cent in every case.

The importance of the arbitrary rule that six times the standard deviations covers 99 per cent of the total frequency is of special significance in the analysis of times series. It will be sufficient here merely to indicate its use. Time series are usually studies involving comparison and require graphic presentation as an aid in analysis. The object of a study may be a comparison

[1] YULE, G. U., "An Introduction to the Theory of Statistics," p. 140.

of the seasonal, secular, or cyclic fluctuations of series over a given period of time. In forecasting, cyclical fluctuations



THE INDEX OF GENERAL BUSINESS CONDITIONS, 1903–14

CHART 50.

become the criteria and these relationships are usually shown by means of charts. To judge relationships the data for the respective series are plotted on the same chart. Association,

however, may be difficult to read unless the various curves lie in close proximity to one another—they must be brought into juxtaposition. Absolute figures will not necessarily give the requisite conformation, since some units may be expressed in millions, while others are in single units. Relative numbers, or percentage figures of the items to their respective totals, reduce the possible range of the various series to 0 to 100 per cent—a considerable range. The standard deviation is a measure which brings diverse series into juxtaposition, since six times its measurement, for any series moderately skewed, irrespective of the unit of measurement, covers 99 per cent of the total frequency. If each fluctuation is expressed as a percentage of its standard deviation, the fluctuations for all the series will *tend* to range between -3 and $+3$ of their standard deviations. This is a very valuable relationship in charting time series, as is indicated in Chart 50.

While this discussion has been more or less of a digression, and the subject will be treated more fully in subsequent chapters, it has been given here in order to emphasize the importance of the standard deviation in statistical studies and the need for a thorough mastery and understanding of this measure.

III. COEFFICIENT OF DISPERSION

In Tables XLIII and XLII the standard and mean deviations were found to be \$8.99 and \$7.38, respectively, while for Table XL the same measures are 6.41 and 5.22 inches respectively. In comparing these figures the question naturally arises: Are the dispersions of Tables XLIII and XLII greater than those for Table XL simply because their absolute numbers are greater? This question is readily answered if the student has a thorough understanding of what the two measures of dispersion are. They both measure "scatteration" from a *standard* (mean, median, or mode). In these illustrations the standard is their respective means. Such standards differ for different series. The wages of boys employed in a factory would certainly have a lower mean wage than the mean wage for the men employed in the same factory. Therefore, the measure of dispersion in *absolute numbers* would tend (depending upon the number of grades of workers) to be smaller for the boys' wages than for the men's. It is conceivable that, among the boys, the differences in their wages from their mean wage would be greater than

among the men, especially if all men were skilled. This indicates at once that in comparing series for dispersion absolute numbers cannot be used because the measurement of the standard from which they were computed may vary considerably. Comparability of the dispersion of series may, however, be obtained by expressing each measure of dispersion as a percentage of the standard from which it was calculated. This gives a comparable measure and is termed the "coefficient of dispersion."

The coefficient of dispersion is expressed by the Pearsonian formula

$$v = 100 \times \frac{\sigma \text{ or } A}{M},$$

in which v = the coefficient.
σ = the standard deviation.
A = the mean deviation.
M = the standard from which measured.

Such a measure is the ratio of the dispersion to its measure of position. One hundred is included in the formula to bring the coefficient to a percentage basis and give size to the coefficient.

Applying this formula to Tables XL and XLII the following results are obtained.

TABLE XLIV

Measures of dispersion	Tables XLII and XLIII		Table XL	
	$v = 100 \times \frac{\sigma \text{ or } A}{M}$	Actual measures	$v = 100 \times \frac{\sigma \text{ or } A}{M}$	Actual measures
Standard deviation...	.....	$8.99	.....	6.48 cm.
Coefficient of standard deviation.....	$\frac{100 \times 8 \cdot 99}{30 \cdot 79}$	29.2%	$\frac{100 \times 6 \cdot 48}{17}$	38.1%
Mean deviation.....	.....	$7.38	.....	5.22
Coefficient mean deviation.....	$\frac{100 \times 7 \cdot 38}{30.79}$	23.9%	$\frac{100 \times 5 \cdot 22}{17}$	30.7%

From Table XLIV it is seen that the series, Table XL, which has the smaller absolute measures of dispersion, when reduced to coefficients, really has the greater dispersion. Therefore in

comparing series for dispersion *the coefficients should be used rather than absolute numbers.*

IV. QUARTILE DEVIATION

90. Definition.—Table XXXIX (page 161) summarizes all the tables used in the chapters devoted to averages. A study of these measures indicates how diverse they are for the different series, even though two of the series have the same mean, $30.79. If the attention is concentrated upon the interquartile range and the medians of the respective series, it will become evident that quartiles differ in their respective differences from the median even in the same series, and also between the respective series.

91. Measures of Quartile Deviation.—The interquartile range, as was brought out in a previous chapter, gives the limits within which may be found the center half of the frequencies of the series. These limits will change according to the relative density of the frequencies above and below the median. In a series where the frequencies, either on one side, or both, cluster around the median, the quartile or quartiles approximate to the measurement of the median, as shown in Table XXXVI (page 151), where the third quartile is only 1.72 inches above the median because of the massing of the frequencies just above the median. The first quartile for the same series is some 3 inches below the median because the frequencies are spread out over more class intervals. It becomes evident that such positions of the quartiles could serve as a measure of dispersion because their positions are moved according to the paucity or density of the frequencies throughout the various class intervals. The mean distance between quartiles may then serve as the desired measure.

In a symmetrical series, as illustrated in Table XL, and Chart 49 (page 164), the quartiles are equidistant from the median, so that the measure of dispersion may become either Median $- Q_1$ or $Q_3 -$ Median, which gives for the series a quartile deviation of 4.5 inches. However, such symmetrical series are the exception rather than the rule in economic or social studies. Therefore one quartile tends to be more remote from its median than the other, according to the nature of the decline of the frequencies throughout the class intervals found on either side of the median. Dispersion measures not direction but scattera-

tion. Therefore the mean of the distances between the quartiles becomes a measure of dispersion, or by formula

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}.$$

By substitution values in Table XXXIV (page 139)

$$\text{Quartile deviation} = \frac{\$36.72 - \$23.71}{2} = \$6.50.$$

92. The Quartile Coefficient of Dispersion.—As in the other measures of dispersion, when comparing series it is necessary to express the measures as relative numbers, relative to the standard from which they are measured. In the quartile deviation the standard is naturally, the sum of the quartiles.

$$\text{Coefficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

By substitution for Table XXXIV

$$\text{Coefficient} = \frac{36.72 - 23.71}{36.72 + 23.72} = 0.215.$$

93. Advantages and Limitations of the Quartile Deviation.—*a. Advantages.*—The advantages of this measure are largely negative rather than positive.

1. This measure may be determined by graphic methods, as indicated in Chart 48 (page 152). Graphs are only approximations, however, so that this advantage should not be given too great weight.

2. It is easy of computation.

3. This is the right measure to use if the problem demands that extreme items be ignored.

4. It can be used to advantage in series having either indeterminate groupings at the extremes, such as "under five" or in series having unequal intervals.

5. The greatest advantage lies in its use as a measure of skewness. This point will be developed in the next chapter.

b. Limitations.—1. A discussion of this measure of dispersion was not included in the paragraphs comparing the other measures of dispersion, because there is a marked difference between this measure and the others. The mean and standard deviations, have a definite standard from which to measure their deviations, whereas the quartile deviation has no definite standard—it arbitrarily indicates the mean of the difference between the quartiles.

2. This measure concerns itself only with the center half of the frequency distribution and does not give any weight to extreme items. Because of this fact it is possible, though highly improbable, that two series may have the same quartile deviation and yet differ in the other measures of dispersion. In that extremes are ignored in its computation the quartile deviation is approximately 60 per cent of the standard deviation.

3. As this measure is not derived from one definite standard it is therefore a rough measure of dispersion and should be used with caution.

V. STATISTICAL RÉSUMÉ

In order to link this chapter on dispersion with those on averages the following table indicates the computation of all the measures of position and dispersion. It is given to afford a review of the various computations and method.

TABLE XLV.—WAGES OF ALL MALE AGRICULTURAL WORKERS HIRED BY THE DAY, FIRST QUARTER OF YEAR 1922

Wages	f	d	$d \times f$	$d^2 \times f$	Cum f
$ 0.00–$ 0.49	0				
0.50– 0.99	338	− 2	− 676	1,352	338
1.00– 1.49	1,096	− 1	−1,096−1,772	1,096	1,434
1.50– 1.99	639	0	Assumed mean		2,073
2.00– 2.49	576	1	576	576	2,649
2.50– 2.99	257	2	514	1,028	2,906
3.00– 3.49	395	3	1,185	3,555	3,301
3.50– 3.99	88	4	352	1,408	3,389
4.00– 4.49	71	5	355	1,755	3,460
4.50– 4.99	6	6	36	216	3,466
5.00– 5.49	18	7	126	882	3,484
5.50– 5.99	4	8	32	256	3,488
6.00– 6.49	9	9	81	729	3,497
6.50– 6.99	0	10	22		
7.00– 7.49	2	11		242	3,499
7.50– 7.99	0	12			
8.00– 8.49	0	13			
8.50– 8.99	0	14			
9.00– 9.49	0	15			
9.50– 9.99	0	16			
10.00– 10.49	1	17	17 3,296	289	3,500
Totals.....	3,500	....	+1,524	13,384	

By permission of National Bureau of Economic Research,

$K = \frac{+1,524}{3,500} = 0.435$ intervals.

Mean $= 1.75 + 0.435 \times 0.5 = \$1.96$

Median item $= \frac{3,501}{2} = 1,750.5$

Median $= 1.50 + \frac{316.5}{639} \times 0.5 = \$1.74$

First quartile $= \frac{3,501}{4} = 875$

$= 1.00 + \frac{537}{1,096} \times 0.5 = \$1.24$

Third quartile item $= \frac{3(3,503)}{4} = 2,625$

Quartile $= 2.00 + \frac{552 \times 0.5}{576} = \$2.47$

Mode $= 1.00 + \frac{639 \times 0.5}{639 + 338} = \$1.32$

Standard deviation $= \sqrt{\frac{13,384}{3,500} + (0.435)^2 \times 0.5}$

$= \sqrt{3.823 - 0.189} \times 0.5 = 0.95$ cents

Coefficient of dispersion $= 100 \times \frac{0.95}{1.96} = 49$ per cent

Mean deviation $= \frac{5,068 + 0.435\ (2,073 - 1,427) \times 0.5}{3,500} = 0.76$ cents

Coefficient of mean deviation $= 100 \times \frac{0.76}{1.96} = 36$ per cent

Quartile deviation $= \frac{2.47 - 1.24}{2} = 0.61$ cents

Coefficient $= 100 \times \frac{2.47 - 1.24}{2.47 \times 1.24} = 33$ per cent

Arbitary Rules

1. $6 \times$ standard deviation = 99 per cent of total frequency.

$6 \times 0.95 = 5.70 = \frac{3,488}{3,500} = 99.4$ per cent

2. Standard deviation = 80 per cent of mean deviation.

$= \frac{0.76}{95} = 80$ per cent.

3. Quartile deviation = 60 per cent of the standard deviation.

$= \frac{0.61}{0.95} = 0.64$ per cent.

References

BOWLEY, A. L.: "Elements of Statistics," p. 136.
DAVENPORT, E.: "Principles of Breeding," chap. XII.
DAVIES, G. R.: "An Introduction to Economic Statistics," pp. 36–42.
ELDERTON, W. P. and E. M.: "Primer of Statistics."
KELLEY, TRUMAN L.: "Statistical Method," pp. 70–93.
KING, W. I.: "Elements of Statistical Method," chap. XIII.
SECRIST, H.: "An Introduction to Statistical Methods."
WEST, CARL: "Introduction to Mathematical Statistics," pp. 45–55.
WHIPPLE, G. G.: "Vital Statistics."

YULE, G. U.: "An Introduction to the Theory of Statistics," chap. VIII.
ZIZEK, FRANZ: "Statistical Averages," part III.

Problems

1. Why is a measure of dispersion sometimes necessary to characterize a series?

2. Why is the range not an adequate measure of dispersion?

3. From what standards may dispersion be measured, and which one is commonly used?

4. How does the mean deviation violate the rules of arithmetic?

5. In what way does the standard deviation rectify the defect in the computation of the mean deviation?

6. What effect has the change in method, referred to in problem 5, on the relative values of the mean and standard deviations?

7. Define the standard deviation.

8. How may the quartiles serve as a measure of dispersion?

9. What is the relationship between the value of the quartile measure of dispersion and the standard deviation? Why?

10. In comparing two series it is found that one has a larger standard deviation than the other. Is it justifiable to assume that there is more scatteration in the former than the latter? If so, why? If not, why not?

11. How would you obtain comparability of dispersion in series?

The following problems are based upon the frequency tables of problem 10, Chap. VII.

12. Compute the mean deviation for the lowest wage series.

13. Compute the standard deviation for each of the three series.

14. Compare the respective standard deviations and then reduce them to coefficients to see if the relationship between them still persists.

15. Compute the quartile measure of dispersion for the lowest wage series.

16. Derive the percentage relationship of the quartile and mean deviation to the standard deviation of the lowest wage series.

17. Observing the percentages above (problem 16) when would you use the respective measures to indicate the degree of scatteration?

CHAPTER XVI

SKEWNESS

94. Introduction.—In previous chapters methods which characterize a frequency series by its "central tendency," and scatteration have been discussed. Central tendencies as measured by the various averages do not always afford sufficient analysis of a series, since they indicate in themselves nothing of the degree of scatteration of the frequencies from the norm. In order to measure scatteration the measures of dispersion were discussed. The question now arises: Are these two measures of position and dispersion sufficient to characterize a series in all its detail?

95. Definition.—Since the mean deviation by computation adds, *irrespective of sign*, the deviations from some standard, it becomes apparent that dispersion measures scatteration in general and has no interest in the direction of dispersion. Direction of dispersion may well be an important measure, since in income studies it may be very important to know how the lower half of the data distribute themselves—whether or not the frequencies for the smaller incomes mass more than those for the larger. A measure may be devised which measures the direction of dispersion and such is termed the "measure of skewness." Skewness, therefore, is another way of characterizing a series.

96. The Need for Such a Measure.—In order to show the need for a measure of skewness, Table XLVI indicates two series which have the same measure of dispersion, standard deviation, yet differ in the direction of dispersion. These series are, naturally, synthetic, being compiled to show the necessity of a measure of skewness. In the cumulative frequency curves or ogives of Table XLVI, as illustrated in Chart 51, it is seen that the Y curve rises abruptly to a high point, from which the curve takes an almost horizontal position. Curve X, on the other hand, rises gradually and ends with a sharp inclination. These two curves are said to be skewed. This perhaps may be

TABLE XLVI.—TABLES ILLUSTRATING HOW TWO SERIES MAY HAVE THE SAME STANDARD DEVIATION BUT DIFFER IN OTHER RESPECTS

Series X						Series Y					
Length in centimeters	f	d	fd	fd^2	f	Length in centimeters	f	d	fd	fd^2	cf
0– 1.9	2	−6	− 12	72	2	2– 3.9	25	−3	− 75	225	25
2– 3.9	3	−5	− 15	75	5	4– 5.9	94	−2	−188	376	119
4– 5.9	6	−4	− 24	96	11	6– 7.9	120	−1	−120	120	239
6– 7.9	15	−3	−45	135	26	8– 9.9	86	0	.....	...	325
8– 9.9	40	−2	− 80	160	66	10–11.9	40	1	40	40	365
10–11.9	86	−1	− 86	86	152	12–13.9	15	2	30	60	380
12–13.9	120	0	.....	...	272	14–15.9	6	3	18	54	386
14–15.9	94	1	94	94	366	16–17.9	3	4	12	48	389
16–17.9	25	2	50	100	391	18–19.9	2	5	10	50	391
Totals.....	391		−118	818	...	.....	391	...	−273	973	

$$K = \frac{-118}{391} = -0.301$$

$$\sigma = \sqrt{8\frac{18}{391} - (0.301)^2}$$

$$= 1.414 \text{ intervals}$$

$$\text{Median} = 12 + \frac{44 \times 2}{120}$$

$$= 12.7 \text{ centimeters}$$

$$\text{First quartile} = 10 + \frac{32 \times 2}{86}$$

$$= 10.7 \text{ centimeters}$$

$$\text{Third quartile} = 14 + \frac{22 \times 2}{94}$$

$$= 14.4 \text{ centimeters}$$

$$\text{Mean} = 13 - 0.301 \times 2$$

$$= 12.4 \text{ centimeters}$$

$$\text{Mode} = 12 + \frac{94 \times 2}{94 + 86}$$

$$= 13.1 \text{ centimeters}$$

$$K = \frac{-273}{391} = -0.698$$

$$\sigma = \sqrt{9\frac{73}{391} - (0.698)^2}$$

$$= 1.414 \text{ intervals}$$

$$\text{Median} = 6 + \frac{77 \times 2}{120}$$

$$= 7.28 \text{ centimeters}$$

$$\text{First quartile} = 4 + \frac{73 \times 2}{94}$$

$$= 5.55 \text{ centimeters}$$

$$\text{Third quartile} = 8 + \frac{55 \times 2}{86}$$

$$= 9.28 \text{ centimeters}$$

$$\text{Mean} = 9 - 0.698 \times 2$$

$$= 7.71 \text{ centimeters}$$

$$\text{Mode} = 6 + \frac{86 \times 2}{86 + 94}$$

$$= 7 \text{ centimeters}$$

better illustrated by Chart 52*a*. Therefore, realizing that there are different positions of concentration, even though dispersion may be the same, it becomes necessary to find some measure for skewness.

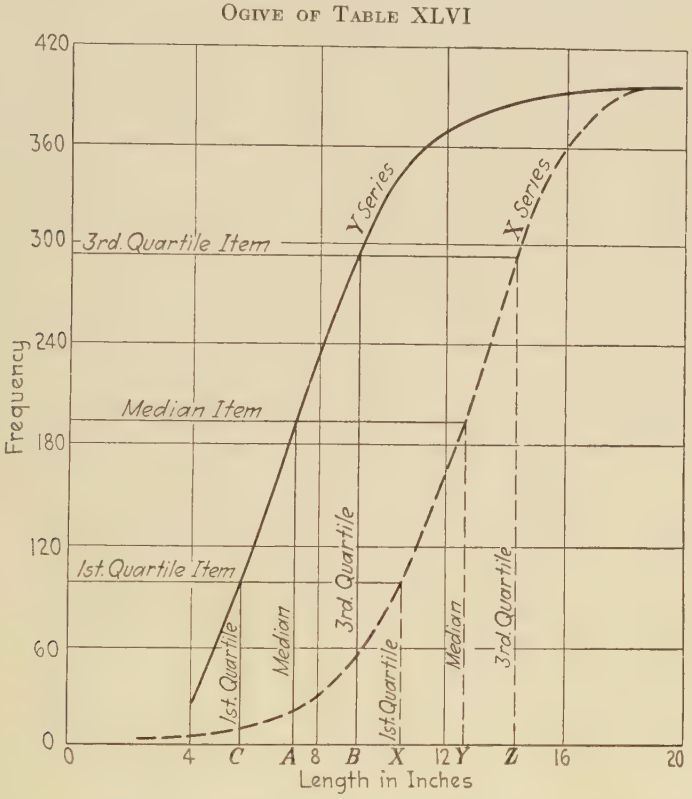


Chart 51.

97. The Measure of Skewness.—*a. Foundation for Measure.*—In turning to Chart 49 (page 164) it will be recalled that it presents a symmetrical series, which implies that the mean, median, and mode, when represented graphically, are conceived as being superimposed one upon the other. In this particular chart the mean, median, and mode all measure 17 inches; likewise, the quartiles are equidistant from the median, each being removed by 4.5 inches. Such a type of series is an exception, since a normal distribution is rarely found. Series are usually distorted in economic studies—the frequencies tend to mass in half of the table more than in the other. Under such condi-

tions, the mean, median, and mode will not have identical measurements, neither will the respective quartiles be equidistant from their median. Such facts, at once, suggest a possible measure of skewness, namely, the relations of the various averages to one another.

b. Mean − Mode Method.—The mean, by definition, divides the sum of the deviations from itself in such a way that their sum is zero. The mode ignores extreme items—its value is only affected by the frequencies immediately around the modal

Frequency Polygons of Table XLVI

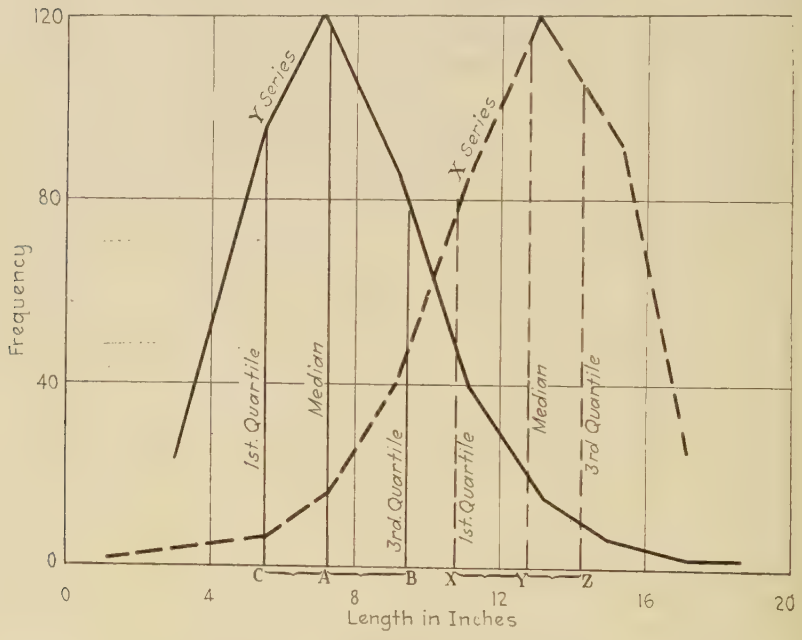


Chart 52.

group. Extreme items, therefore, tend to pull the mean towards them, while they have no effect upon the mode. This fact at once suggests a possible measure, the difference between the mean and mode (Mean − Mode).

As already stated, in a symmetrical series the mean, median, and mode would be identical, but in ordinary series where distributions are asymmetrical, the mean would be *larger* than the mode if there were more extreme items in the upper classes of the table than in the lower. This condition is evident in

Table XLVI, series Y, in which the mode is 7 and the mean 7.71 centimeters. This series has a positive skew. Chart 52 indicates how the frequencies for series Y tail off gradually in the upper classes. The meaning of the term "positive skew" should be apparent from the chart and table. In a wage study the existence of a positive skew would be a healthy sign, for it would mean that there was a tendency for greater dispersion in wages above the mean age than below it.

If there were more extreme items in the lower half of the table, in contrast to the upper, the mean would be pulled towards the lower classes, the mode being unaffected. Series X of Table XLVI illustrates this, the mean being 12.4 and the mode 13.1 centimeters. The skew in this particular case is negative, indicating that there is greater scatteration in the lower half of the table.

Skewness may be measured by the formula

$$\text{Mean} - \text{Mode}.$$

By substitution of values for Table XLVI

$$\text{Skew of series } X = 12.4 - 13.1 = -0.7.$$
$$\text{Skew of series } Y = 7.71 - 7. \quad = +0.7.$$

c. Quartile Measure.—From the discussion of the quartile deviation it will be recalled that the quartiles deviate from their median according to the relative density of the frequencies about the median. Series tend to be irregular in this density and one quartile is further removed from its median than the other. This fact at once proves the existence of skewness or asymmetry, because quartiles are equidistant from their median in symmetrical series. A measure of asymmetry or skewness may therefore be developed from the respective positions of the quartiles of a series. The more distant quartile suggests a tailing off of the frequencies, thus giving a skewed conformation to the polygon as illustrated in Chart 52.

The position of skewness may, therefore, be found by comparing the respective distances of the quartiles from their median: or by formula

$$(Q_3 - \textit{Median}) - (\textit{Median} - Q_1).$$

By simplifying

$$\text{Skewness} = Q_3 + Q_1 - 2\,\textit{Median}.$$

By substitution,

Series X, Table XLVI

Skewness $= 14.4+10.7-2(12.7) = -0.3$ centimeters

Series Y

Skewness $= 9.28+5.55-2(7.28) = +0.3$ centimeters.

The computation above and Chart 52 indicate that if the distance YZ, Chart 52, between the third quartile and the median is smaller than that from the first quartile, YX the skew is negative. Conversely, if the third quartile is more remote (AB, Chart 52) from the median than the first quartile AC, the skew is positive.

Positive skewness, therefore, indicates a massing of frequencies below the median with little scatteration, while negative skewness indicates the concentration above the same measure.

98. Coefficients of Skewness.—It is very often desirable to compare series. This implies the necessity of expressing the measures in comparable terms. As in the treatment of coefficients of dispersion, the measures must be divided by the standards from which they were computed.

a. Mean-Mode.—Naturally, since skewness measures the direction of scatteration, the logical standard with which to measure skewness as a relative would be the standard deviation as used in the Pearsonian formula. The coefficient is

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}.$$

By substitution Table XLVI

$$\text{Series } X = \frac{12.4 - 13.1}{2.8} = -0.25$$

$$\text{Series } Y = \frac{7.71 - 7}{2.8} = +0.25$$

b. Quartile Measure.—Since this is a quartile measure the logical standard to use in the derivation of the coefficient would be the quartile deviation $\frac{Q_3 - Q_1}{2}$.

The coefficient thus becomes

$$\frac{Q_3 + Q_1 - 2\,\text{Median}}{\frac{Q_3 - Q_1}{2}}.$$

By simplifying

$$\text{Coefficient of skewness} = \frac{2(Q_3 + Q_1 - 2\,\text{Median})}{Q_3 - Q_1}.$$

By substitution for Table XLVI

$$\text{Series } X = \frac{2(14.4 + 10.7 - 12.7\times 2)}{14.4 - 10.7} = 0.16.$$

$$\text{Series } Y = \frac{2(9.28 + 5.55 - 2 \times 7.28)}{9.28 - 5.55} = 0.16.$$

These figures in comparison to those computed by the mean − mode method again emphasize the defect of quartile measures, since the values are much smaller because the extreme items are not included in the computation.

99. Summary.—From this chapter it should be evident that it is not sufficient in all cases to characterize a series by measures of dispersion or position. Direction may be equally as important as dispersion. Therefore skewness is an important measure. Of the two measures of skewness the quartile measure should be used with caution, since it is a measure derived from only half the data of a series. Comparability in skewness is achieved by computing coefficients.

References

DAVIES, G. R.: "Introduction to Economic Statistics," p. 48.
KELLEY, TRUMAN L.: "Statistical Method," p. 77.
KING, W. I.: "Elements of Statistical Method," chap. XIV.
SECRIST, H.: "An Introduction to Statistical Methods."
WEST, CARL: "Introduction to Mathematical Statistics," pp. 57–58.
YULE, G. U.: "An Introduction to the Theory of Statistics," pp. 149–153.

Problems

1. How does a measure of skewness supplement the information given by the measure of dispersion?

2. Why is the mode used, in one of the methods, as the standard from which to measure skewness?

3. Why do the differences between the quartiles serve as base in the derivation of the quartile measure of skewness?

4. What is meant by the terms "positive" and "negative" skew?

5. By glancing at the table in problem 6, Chap. VIII would you say that both series, columns A and C, have the same skew? Explain your answer.

The following problems apply to problem 10, Chap. VII.

6. Compute the measures of skewness, by the various methods, for all three series.

7. Derive the coefficient of skewness for each.

8. From the coefficient above (problem 7), which of the three series indicates the best conditions for wages in general?

CHAPTER XVII

CORRELATION

100. Résumé of Preceding Chapters.—The preceding chapters have discussed the methods of attacking a statistical problem. These methods may be summarized briefly as follows:

1. Having the problem definitely in mind.
2. Selection of and rigid definition of the unit of measurement.
3. Collection of material
   - *a.* Primary data
     - Enumeration
     - Estimation
   - *b.* Secondary data
     - Analysis of unit
     - Bias
     - Homogeneity
4. Classification and tabulation
   - Time series
   - Spatial
   - Component
     - Stub and captions
   - Frequency
     - Class interval
     - Class limits
5. Aids in presentation
   - Diagrammatic—bar charts, etc.
   - Graphic
     - Polygons or curves
     - Ratio charts
     - Histograms
6. Steps in analysis
   - *a.* Measures of position
   - *b.* Measures of dispersion
   - *c.* Measures of skewness

101. Correlation and Regression.—All chapters in Part II have been devoted to the study of problems which arise in frequency series. Analysis of such tables usually implies comparison, and yet, as already indicated, the various steps in analysis have their limitations. Averages give position only; dispersion, scatteration without direction; and skewness direction without position or degree of dispersion. Therefore, to characterize a series properly all of these measures should be computed. But the question arises: Does such a composite picture suffice in all cases? It is clear that each of the above measures is but a picture of a section of one series and not of the entire series; relationships have been isolated within a series and not shown *between series*. They do not show how each and every item in

one series relates itself, in general, to the items of the second series. Hence, sometimes it is necessary to find out if there is association between series and to derive some measure which will bring out not merely a measure of position, dispersion, or skewness of two series, but rather relationships between all items. The study of this relationship, the association of the items of one series with those of another, involves correlation, while the study of the degree of variation within the related series is that of regression.

102. History of Coefficient of Correlation.—Sir Francis Galton in 1888, in his study of "Family Likeness in Stature," endeavored to trace the relationship of the heights of parents with those of offspring. In this study he found that the offspring "regressed" or stepped back towards type; that is, if the mean height of certain parents was in excess of the mean height of adults in general, *e.g.*, to the extent of 2 inches, then the height of the children of such parents would, generally, at maturity be greater than the average adult, but would deviate from the mean height of adults to a lesser extent than the parents, *e.g.*, the deviation being about 1 inch. The children, therefore, tended to revert back to type. Galton called this association which he found between the heights of parents and offspring "regression." With the aid of the mathematician J. D. Dickson, a mathematical device was found for measuring the association between two variables, and was given in an article on "Correlation and Their Measurement," in the *Proceedings of the Royal Society, London,* 1886. The name of this measure was changed to "coefficient of correlation" by Professor Edgeworth in his paper on "Correlated Averages," written for the Royal Society. The development of the coefficient of correlation, as now used, may be attributed to the efforts of Karl Pearson, Professor Edgeworth, and G. Undy Yule. Some writers attribute the first development of correlation to Bravais, but, as one writer has pointed out, such claims are unwarranted. "Galton alone seems deserving of being called the father of correlation."[1]

103. Definition.—From the above superficial development of the history of correlation it must be apparent that correlation is *not causation,* for it is not a law that all tall parents have tall children, but, in general, there is a possibility of such being the case. In other words, correlation is a study of *possibilities*

[1] KELLEY, TRUMAN L., "Statistical Method," p. 152.

of association. It is well known that, in general, a person's weight increases with height, that is, there is association, a tall person usually weighing more than a short one. This relationship is a positive one, height and weight increasing together. Height cannot be said to cause weight or *vice versa*, for it is also well known that weight also increases with age. *Therefore correlation is a measure of association and not causation.* It is true that there may be some causal connection between two series, but in most economic phenomena it is impossible to isolate cause or to determine effect. The student is referred to the articles written by W. I. King, and E. E. Day, in the *American Statistical Publication* for December, 1917 and June, 1918, in which the above topic is discussed. Correlation has been defined as follows:

> When two quantities are so related that the fluctuations in one are in sympathy with the fluctuations of the other, so that an increase or decrease of one is found in connection with an increase or decrease (or inversely) of the others, and the greater the magnitude of the change of the one the greater the magnitude of the change in the other, the quantities are said to be *correlated.*[1]

Another brief definition emphasizing the idea of association in contrast to causation is that of W. Brown.

> Correlation may be briefly defined as the tendency toward concomitant variation, and the so-called correlation coefficient is simply a measure of such tendencies more or less adequate according to the circumstance of the case.[2]

The importance of the probability of such association can perhaps best be indicated by referring to the well-known policy of insurance companies. Insurance companies through actuarial methods have studied the death rate per thousand of persons in normal health, according to age, sex, and occupation. This rate over a period of time has afforded the basis of the life-expectancy table, from which are derived the premiums to be paid by policy holders. Life expectancy, naturally, diminishes with age; that is, there is a negative relationship. Therefore older people, when taking out insurance, have to pay a higher premium than younger people who are of normal health and working in an occupation with no greater risk. All kinds of

[1] Bowley, A. L., "Elements of Statistics," p. 316.

[2] Brown, W., "The Essentials of Mental Measurement," p. 42.

insurance—fire, burglary, etc.—are based in a similar fashion on past associations of various phenomena.

Much social legislation is written in an attempt to avoid deleterious associations, such as unsanitary conditions and the health of the worker; accidents and the lack of safety appliances; child labor and backward communities; immigration and the lowering of the standard of living. Some of this type of legislation has not the desired effect, because the association presumed to exist was actually non-existent—not because the evils were imaginary, but rather that the legislation was based *upon casual observance* of an accidental resemblance between the two phenomena chosen. Association may, therefore, be accidental and should be tested for the degree of association. This may be done by computing the correlation coefficient. This, however, is not infallible, but it is more effective than *casual* observance.

The importance of correlation will be further emphasized in Part III, which is devoted to the study of time series. It will be sufficient to state here that, by means of correlation of time series, forecasters are being derived with the hope that these may smooth out to a certain extent the curves of the industrial cycle.

The first step, then, is to develop some means of measuring the degree of association existing between two or more series. It must be a method whereby one may measure whether the increase in one series *e.g.*, weight, is associated with an increase in the other, height. A measure which shows such relationship would indicate *positive correlation*, which means that an increase in one series is represented by an increase in the related series. If, on the other hand, an increase in one series, age, is associated with a decrease in the related series, life expectancy, then the measure is that of *negative correlation.*

104. The Correlation Table.—The first step in a correlation study is to bring the related items together. For frequency series this is accomplished by means of a correlation table, which shows the paired items in juxtaposition. Such a table is really the condensation of two tables into one; the class intervals of one running horizontally across the page, while that of the other reads down, at the left. The related frequencies are then recorded in the proper intersections of rows and columns, thus showing, *not measuring*, the relationship. This may best be illustrated by Table XLVII.

TABLE XLVII.—CORRELATION TABLE OF HEIGHT AND WEIGHT OF 250 STUDENTS

(X) Height in inches	Weight in pounds (Y)								Total
	100 to 109	110 to 119	120 to 129	130 to 139	140 to 149	150 to 159	160 to 169	170 to 179	
60–61.9	1	1	10	..	..	..	..	..	12
62–63.9	..	..	9	18	3	..	..	..	30
64–65.9	..	1	..	18	25	9	3	..	56
66–67.9	..	..	3	19	40	11	..	1	74
68–69.9	..	..	1	..	25	29	1	3	59
70–71.9	..	..	1	1	..	7	4	6	19
Total.....	1	2	24	56	93	56	8	10	250

It is evident from this table that the correlation to be measured is between the height and the weight of 250 students. From the chapter on Tabulation the method of recording the data in the table is evident. The frequency 40 marked □ was derived by placing a mark in this intersection when the weight of a student was found to be within the limits of 139.5 to 149.5 pounds and at the same time measured in height between 66 to 67.9 inches; that is, a person of a height of 67 inches and weighing 143 pounds would be recorded in this group. Of the 250 students, 40 were found to vary within the two limits of height and weight of this class. The total, 12, for the first row indicates that twelve persons varied in height from 60 to 61.9 inches, while their weights differed considerably, as is shown by the allocation of the 12.

From correlation tables some idea of association may be derived from the conformation of the frequencies. The totals in the above table synchronize, more or less, the frequencies massing in the center half of the table. There are few frequencies for the smaller weights or heights, indicating a positive correlation. The degree of association may be roughly estimated by the scatteration of the frequencies about the diagonal across the table. In perfect association, one would expect the frequencies to be compressed into a small area about the diagonal, because, as the value of one of the series increased, that of the other would vary directly, or inversely in a definite ratio. Table

XLVII shows no such marked concentration, as the frequencies scatter and indicate a moderate degree of association ($r = +0.65$) Table XLVIII indicates more effectively the approach to perfect association or correlation.

TABLE XLVIII.—CORRELATION STUDY (BY SAMPLES) OF VOTES RECORDED IN CUYAHOGA COUNTY, OHIO, JUNE, 1914

Percentage of vote on suffrage cast in favor of suffrage

Percentage of vote on prohibition cast in favor of prohibition	5–9.9	10–14.9	15–19.9	20–24.9	25–29.9	30–34.9	35–39.9	40–44.9	45–49.9	50–54.9	55–59.9	60–64.9	
0– 4.9	1												1
5– 9.9	6	4											10
10–14.9	2	2	3	1	2								10
15–19.9		2	3	3	1								9
20–24.9			3	6	3								12
25–29.9				1	4	1		1					7
30–34.9		1			2	1	4	2					10
35–39.9						2		4	1				7
40–44.9							1	5	1	2			9
45–49.9								1	2	6	1		10
50–54.9								1	2	4	4	1	12
55–59.9									1		1	1	3
	9	9	9	11	12	4	5	14	7	12	6	2	100

Gehlke, C. E., "Correlation between Vote for Suffrage and Vote on Liquor Question," Amer. Stat. Pub., March, 1917.

In this table the frequencies mass decidedly around the diagonal sloping to the right. Association or correlation is therefore positive and quite marked (+0.87).

TABLE XLIX.—PERCENTAGE OF VOTE ON SUFFRAGE CAST IN FAVOR OF SUFFRAGE

Percentage of vote on license cast in favor of license	0–4.9	5–9.9	10–14.9	15–19.9	20–24.9	25–29.9	30–34.9	35–39.9	40–44.9	45–49.9	50–54.9	55–59.9	60–64.9	65–69.9	70–74.9	
55–59.9			1									1				2
60–64.9											1	2				3
65–69.9									3	1	3		1	2		10
70–74.9			1			3			2	3	1	1		1	1	13
75–79.9						3	1	6	4	5	1	1				21
80–84.9		1		2	1	4	3	3	3	2	2					21
85–89.9		2	4	5	3	3		2	1	2						22
90–94.9			2		2	1	1			1						7
95–99.9	1															1
	1	3	8	7	6	14	5	11	13	14	8	5	1	3	1	100

Table XLIX, a part of the same study, shows a definite concentration, but also shows an inverse variation, the frequencies massing around the diagonal drawn up and from the left. This indicates a *high* value in one of the two related series is associated with a *low value in the other*, thus giving negative correlation ($r = -0.59$).

Correlation is a function which expresses the relation which on the average exists between the two variables involved. It tells how the data scatter around the lines of regression.

The above discussion of the various conformations may appear to be a digression. However, by a study of these tables a better idea of what correlation is should be obtained and, furthermore, it should aid in the understanding of the methods used in computing the measure.

105. How to Measure Correlation by Means of Standard Deviation Measures.—The problem is to devise some method whereby it may be determined whether or not high values in the one series associate themselves with high values in the other—in other words, to measure positive correlation, or high values in one are associated with low values in the other (negative correlation); and also to indicate to what degree this association is to be found.

It was suggested in the previous paragraphs that correlation is more or less indicated by the conformation of the frequencies about the diagonals of the correlation table. This, at once, suggests a possible measure, for if the center of gravity of the table is found and the respective deviations for each of the paired items measured from this point, by comparing these variations the relationship of the series may be found. The center of gravity of a table is graphically the intersection of the two means of the series (see Chart 53).

Table L.—Length and Breadth of X

Length in centimeters	Breadth in centimeters
20	8
24	10
28	12
32	16
36	18

In order to illustrate and develop the method of computation a very simple series is given in Table L.

This table is made into a correlation table, for illustrative purposes only. It should be remembered that the figures are actual measurements and there are no class limits.

TABLE LI

X series Length in centimeters	*Y* series Breadth in centimeters				
	8	10	12	14	16
20	1				
24	..	1			
28	..	..	1		
32	..	..	..	1	
36	..	..	..	..	1

In glancing at this reconstructed table (Table LI) it becomes apparent that if there is any correlation it must be positive, since with an increase in length the breadth increases in magnitude. It would, naturally, be expected, if the association is perfect, that the variations in the one series would show equal variations in the other. These variations, however, cannot be compared in absolute numbers, since in absolute numbers length increases more than breadth. Therefore, when equal variations are spoken of, equal percentage of change should really be meant. The problem of measuring correlation therefore reduces itself to the derivation of a standard from which to measure percentage change.

Chart 52*a* (page 212) illustrates the data of Table L. The related items were plotted at the intersection of their respective abscissa and ordinate. A cross is placed at the intersection of the abscissa corresponding to length 20 and the ordinate of its related breadth 8. The center of gravity of the series is shown at 0, where the line of mean breadth 12 intersects that of mean length 28. Since the sum of the deviations from the mean equals zero, the related variations may be measured from this point. The related pair of length 20 and breadth 8 deviate −8 and −4 from the center of gravity.

It should be recalled that two series may have identical means and yet differ in dispersion. Therefore to reduce the deviations to a comparable basis it is best, for this preliminary method, to express the various deviations from their means in terms of the respective standard deviations. In the illustration length 20 deviates from its mean -8 and, since the standard deviations for this series is 5.66, this deviation equals $\frac{-8}{5.66}$, or -1.41 in terms of the standard deviation, as shown in column C, Table LII. Such values when computed for the two series should show if the series disperse regularly in the same degree and direction. For perfect correlation the paired items must synchronize in their degree of dispersion, although they may vary in direction, varying directly for positive and inversely for negative correlation. Table LII indicates the method of computation.

TABLE LII

X series Length in centimeters					Y series Breadth in centimeters				XY	
Size	Dev.	Dev.^{2}	$\frac{\text{Dev.}}{\text{St. Dev. } X}$	X^2	Size	Dev.	Dev.^{2}	$\frac{\text{Dev.}}{\text{St. Dev. } Y}$	XY in St. Dev.	XY from mean
	A	B	C	D	E	F	G	H	I	J
20	-8	64	-1.41	1.98	8	-4	16	-1.41	1.98	32
24	-4	16	-0.70	0.49	10	-2	4	-0.70	0.49	8
28	0	...	.....	....	12	0				
32	4	16	0.70	0.49	14	2	4	0.70	0.49	8
36	8	64	1.41	1.98	16	4	16	1.41	1.98	31
140	...	160	.....	5.94	60	...	40	.....	5.94	80

Mean $= \frac{140}{5} = 28$ centimeters
Standard deviation $= \sqrt{\frac{160}{5}} = 5.66$

Mean $= \frac{60}{5} = 12$ centimeters
Standard deviation $= \sqrt{\frac{40}{5}} = 2.83$

The computations of the columns of this table are familiar, except columns D and I. These two form the basis of the measure of association. Each deviation from the mean has been expressed in terms of its standard deviation, as shown in columns C and H. It has been stated that for perfect correlation the two series must disperse regularly and in the same degree. Therefore, if the sum, Σxy, of the product of each deviation in X series and its related deviation in the Y series, column I (the deviations being expressed as values of their standard

deviations), be divided by the sum of the deviations squared of the X series ΣX^2, column D, a measure of correlation is found so that:

$$r = \frac{\Sigma xy}{\Sigma x^2},$$

in which r = the coefficient of correlation.

x = deviations, in terms of the standard deviation, of the items in the X series.

y = related deviations in the Y series.

By substitution for Table LII.

$$r = \frac{5.94}{5.94} = 1.$$

The reason for dividing by the sum of the deviations squared of one of the series is no doubt self-evident, but it should be emphasized that perfect correlation implies perfect association in the degree of dispersion of the two series. Therefore, if the series are perfectly correlated Σxy should equal Σx^2 and the correlation coefficient is 1, since X equals Y. To determine if X equals Y, the X's should be multiplied by their respective Y's, summated, and divided by the summation of X's squared. For corroboration of this reasoning, it should be observed that columns C and H are identical, thus making Σxy equal Σx^2, likewise shown in column I.

On understanding the above facts it should be recognized that the coefficient of correlation can never be *larger than* 1, and is a measure of the average association. Furthermore, the method of calculation should indicate that a coefficient of correlation, unless association is perfect, cannot, in itself, be used as a forecaster of changes in the series; that is, given a change in the X series the change in the related Y cannot be estimated, such ratio of variation may only be determined by the lines of regression.

It is unnecessary to work out a negative coefficient. For an understanding of such a measure all that is required is to reverse the measurements of the Y series, having the greatest breadth related to the smallest length. By so doing, the Y deviation will be plus when the X is minus and *vice versa*, the sum of these products will, therefore, be minus, giving perfect but negative correlation.

The above formula, though frequently used in time series, is rarely used in frequency series, since it involves an extra and

needless step in computation. However, it has been given in order to make the coefficient more readily understood than if the Pearsonian formula had been given. The extra computation involved is that of expressing all deviations from the mean as relatives to the standard deviation, columns C and H, Table LII.

106. Computation by Actual Deviations from the Mean.—The coefficient of correlation may be obtained from the summation of the product of the respective deviations *from the mean.* When all the measurements are in terms of the standard deviation, the formula is

$$r = \frac{\Sigma xy}{\Sigma x^2}. \tag{1}$$

If it is desired to have the numerator of the formula in terms of deviations from the means, and thus avoid a needless computation, it may be adjusted by adding the product of the respective standard deviations to the denominator of the formula ($\sigma_x\sigma_y$). This product should be added because correlation cannot ignore dispersion and this addition reduces the units to terms of the standard deviations. By substitution Eq. (1) becomes

$$r = \frac{\Sigma xy}{\Sigma x^2\sigma_x\sigma_y}.$$

Substitution, however, must be made for Σx^2, since it, in itself, is computed from deviations in terms of their standard deviation, as shown in column D, Table LII. By consulting this table it should be apparent that

$$\Sigma X^2 = \Sigma\left(\frac{d}{\sqrt{\frac{d^2}{N}}}\right)^2, \text{ so that } \Sigma X^2 = \Sigma N.$$

By substitution of this value in the above formula the Pearsonian formula is derived, in which

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}. \tag{2}$$

x = deviations measured from the *mean* of X series.
y = deviations measured from the *mean* of Y series.
σ_x = standard deviation of X series.
σ_y = standard deviation of Y series.

Application of Method to Table LII.—In this table Σxy becomes the sum of the products of the deviations of the paired items. As shown in columns A, F, and J, this sum equals 80, while

$\sigma_x = 5.66$, $\sigma_y = 2.83$, and N equals 5. By substitution of these values in the formula,

$$r = \frac{80}{5 \times 5.66 \times 2.83} = 1.$$

The result is identical with that of Eq. (1).

Such examples as those already given are rarely found in the measurement of variable phenomena. The items in most frequency tables involve multiple frequencies and the need for class intervals, and thus give rise to correlation tables.

In correlation tables the frequencies and the class intervals make the computation of Σxy a little more involved. The other computations are similar to those already given. The standard deviation is computed for each of the series. The computation of the standard deviation should not afford any difficulty, although the frequencies intervening between the totals seemingly confuses the student. The totals for the Y series or weight in Table LIII are found at the bottom of the table, while those for the X series are at the extreme right. The summation of the class frequencies for the respective series must be equal. The total is found at the lower right-hand corner. This indicates the number of paired items or N.

The standard deviation for the X series may be computed at the right of the table itself, as in Table LIII.

TABLE LIII.—CORRELATION STUDY OF HEIGHTS AND WEIGHTS OF 250 STUDENTS

X Height in inches	Weight in pounds								Total	Computation of mean and dispersion of series X		
	100–109.9	110–119	120–129	130–139	140–149	150–159	160–169	170–179		d	$d \times f$	d^2f
60–61.9	1	1	10	...	...	...	...	...	12	−3	−36	108
62–63.9	..	..	9	18	3	...	...	...	30	−2	−60	120
64–65.9	..	1	...	18	25	9	3	...	56	−1	−56	56
66–67.9	..	..	3	19	40	11	...	1	74	0		
68–69.9	..	..	1	...	25	29	1	3	59	1	59	59
70–71.9	..	..	1	1	...	7	4	6	19	2	38	76
Total.....	1	2	24	56	93	56	8	10	250	...	−55	419

It is better to compute the standard deviation for the Y series separately as below.

STANDARD DEVIATION OF WEIGHT OR Y SERIES

Measures of Y

Weight in pounds	f	d	$f \times d$	$f \times d^2$
105	1	−4	− 4	16
115	2	−3	− 6	18
125	24	−2	−48	96
135	56	−1	−56	56
145	93	0		
155	56	1	56	56
165	8	2	16	32
175	10	3	30	90
Total	250	...	−12	364

$$K = \frac{-12}{250} = 0.048$$

$$\text{Standard deviation} = \sqrt{364\tfrac{4}{250} - (048)^2} \times 10$$
$$= 12.05 \text{ pounds}$$

The measures of X and Y series as derived from the tables are:

Measures of X

$K = -55\tfrac{5}{250} = -0.22$. Mean $= 67 - 2 \times 0.22 = 66.5$ inches.

Standard deviation $= \sqrt{419\tfrac{9}{250} - (0.22)^2} \times 2 = 2.552$ inches.

Measures of Y

$K = 12\tfrac{2}{250} = -0.48$. Mean $= 145 - 0.5 = 144.5$ pounds.

Standard deviation $= \sqrt{364\tfrac{4}{250} - (0.048)^2} \times 10 = 12.05$ pounds.

With the determination of the above values the only unknown value for the computation of the coefficient of correlation is Σxy.

To illustrate the procedure for the deriving of Σxy the class interval of 60 to 61.9 series X is chosen. Within this group are found twelve items, all related to the class interval of height but having certain variations with regard to weight. Correlation requires that measurements be made from the respective means. The mean for the X series is found to be 66.5 inches, which yields a deviation for the class interval 60 to 61.9 of −5.5, since 61 is the midpoint of the class. All the twelve items in this height grouping therefore differ from the mean of

the X series by -5.5, because the twelve persons are all of an approximate height and so grouped as a class. Differences, however, should be computed from the mean of the Y series in order that the related x and y deviations may be multiplied together. The mean for the Y series is 145.5 pounds, from which the deviations of the items may be found. The table indicates one person between the heights of 60 and 61.9 inches to have a weight within the limits of 100 to 109 pounds. The y deviation for this frequency must be 144.5–105 (midpoint of class), which is -39.5. Therefore for interval 61 of the X series and 110 for the Y the relationship when measured in deviations equals $1 \times -5.5 \times -39.5$ for xy. By the same procedure all the relationships of the twelve items for the interval 60 to 61.9 may be found. Of course, it is evident if a frequency occurs greater than 1, as in group 120 by 129, which has a frequency of 10, the actual relationship xy should be multiplied by the frequency, because each frequency deviates -5.5 and -19.5. By this method the sum of the deviations for the row 60 to 61.9 is found to be 1,452.

DERIVATION OF Σxy FOR ROW 60 TO 61.9 INCHES OF TABLE LIII

X series	Y series	f	Deviations from means		Product $xy \times f$
Height in inches	Height in pounds		x	y	
60–61.9	100–109	1	−5.5	−39.5	217.25
60–61.9	110–119	1	−5.5	−29.5	162.25
60–61.9	120–129	10	−5.5	−19.5	1,072.25
Total.....	.....	..	.....	−88.5	+1,452

This laborious method should be adopted for every row of the table, as shown in the following table, which is worked out in the same way. In this method care should be exercised in recognizing the signs, since the x deviations may be plus while their related y be minus and thus yield a minus product. By a pursual of this process throughout the table Σxy is found to be $+5{,}087.5$. By substituting all the known values in the formula the coefficient for this table is found to be

$$r = \frac{5{,}087.5}{250 \times 2.552 \times 12.05} = 0.661.$$

TABLE LIV.—COMPUTATION OF COEFFICIENT OF CORRELATION BY LONG METHOD

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$$

Mean of X = 66.5 inches
Standard deviation = 2.552
Mean of y = 144.5 pounds
Standard deviation = 12.05

Row	f	Deviations from means		Product $xy \times f$	
		x	y	+	−
60-62	1	−5.5	−39.5		
	1	−5.5	−29.5		
	10	−5.5	−19.5	1,452	
62-64	9	−3.5	−19.5		
	18	−3.5	− 9.5		
	3	−3.5	+ 0.5	1,207.5	
64-66	1	−1.5	−29.5		
	18	−1.5	− 9.5		
	25	−1.5	+ 0.5		
	9	−1.5	+10.5		
	3	−1.5	+20.5	48	
66-68	3	+0.5	−19.5		
	19	+0.5	− 9.5		
	40	+0.5	+ 0.5		
	11	+0.5	+10.5		
	1	+0.5	+30.5	.....	36.5
68-70	1	+2.5	−19.5		
	25	+2.5	+ 0.5		
	29	+2.5	+10.5		
	1	+2.5	+20.5		
	3	+2.5	+30.5	1,023.75	
70-72	1	+4.5	−19.5		
	1	+4.5	− 9.5		
	7	+4.5	+10.5		
	4	+4.5	+20.5		
	6	+4.5	+30.5	1,392.75	
Totals.....	250	.....	.....	5,124	−36.5

$$\text{Coefficient of correlation} = \frac{5,087.5}{250 \times 2.552 \times 12.05} = 0.661.$$

The foregoing method is time consuming and liable to error in the endless multiplications involved in obtaining Σxy. It is similar in this respect to the long method in the derivation of the standard deviation, and the defect may be surmounted in a similar manner, by resort to the interval method of computation.

CHAPTER XVIII

CORRELATION (*Continued*)

107. The Interval Method of Computation.—The Pearsonian formula $r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$ necessitates that all the calculations be made in absolute units. Naturally, the relationship of the formula is not changed by making equal substitutions in both numerator and denominator; the formula may, therefore, be written $\frac{\Sigma xy}{N\sigma_x\sigma_y}$ when *all* the measurements are made in *class intervals*, and not in absolute numbers. In order to differentiate this formula from that of Eq. (2), the lettering of the numerator is changed from Σxy to Σab, a indicating deviations, *in class intervals, from the mean of the X series*, while (b) measures respective deviations in the other, or Y series. The formula by interval method then becomes

$$r = \frac{\Sigma ab}{N\sigma_x\sigma_y}.$$

But measuring class intervals from the *actual mean* would be just as laborious as the other method. Therefore, it is advisable to adjust the formula in order that the deviations may be measured in class intervals *from an assumed mean.*

Arbitrary means must be assumed for both series. These are shown in Table LV by the broken lines which inclose the respective assumed mean groups. The assumed mean for the Y series, or weight, is 145 pounds, while that for the X series, or height, is 67 inches. It should be observed that the broken lines intersect at frequency 40, indicated by □. This, of course, is the *assumed* center of gravity of the two series from which the deviations in intervals should be calculated.

TABLE LV.—CORRELATION STUDY OF HEIGHT AND WEIGHT OF 250 STUDENTS

(X) Height in inches	Weight in pounds (Y)								Total	Computation of mean and standard deviation series X		
	100 to 109	110 to 119	120 to 129	130 to 139	140 to 149	150 to 159	160 to 169	170 to 179		d	df	d^2f
60–61.9	12 1	9 1	6 10						12	−3	−36	108
62–63.9			4 9	2 18	0 23				30	−2	−60	120
64–65.9		3 1	2	1 18	0 25	−1 9	−2 3		56	−1	−56	56
66–67.9			0 3	0 19	40	0 11		0 1	74	0		
68–69.9			−2 1	−1	0 25	1 29	2 1	3 3	59	1	59	59
70–71.9			−4 1	−2 1		2 7	4 4	6 6	19	2	38	76
Total.....	1	2	24	56	93	56	8	10	250	...	−55	419

The small figures in the above table indicate the frequencies, while the large ones measure deviations in intervals. Since correlation is derived by obtaining the sum of the products of the related items, it is at once apparent that the product (ab) of the deviations within the broken lines must be zero. This may be illustrated by the frequency 25 for the classes of 140 to 149 pounds and 64 to 65.9 inches in height. The a deviation for height is -1, since the group 64 to 65.9 inches is one class interval, in size, below the assumed mean height. The b deviation for weight is 0, since the frequency is in the assumed mean weight group. The product, therefore, of ab, equals zero (-1×0). By a similar process all products of the deviations within the broken lines equal zero. These frequency figures therefore may be ignored in the derivation of the value for Σab.

The method of the determination of the deviations, as represented by the larger figures, may be illustrated by the first frequency in the table which has class limits of 100 to 110 and 60 to 61.9. Since the height of this group, 60 to 61.9 inches, is three class intervals below the assumed mean group, a therefore equals -3. Likewise the weight group, 100 to 110 pounds, is four intervals below the assumed mean weight class, b therefore equals -4, so that ab equals $-3 \times -4 = +12$. This is the

deviation for this group. Some intervals, it should be observed, are minus in sign, group 70 to 71 inches in height with its related weight group 120 to 129 pounds has product *ab* equal to −4. This product arises because the height class is two intervals above its assumed mean ($a = +2$), while the weight is two intervals below its assumed mean ($b = -2$). *ab* therefore equals -2×2, or −4. Therefore, the derivation of the interval deviations is relatively simple; the quadrants *above* (pace) *and to the left* and *below and to the right* are positive while those *above and to the right* and *below and to the left* are negative.

The sum Σab may now be obtained by adding all the *frequencies* which have the *same deviation*, taking care to notice signs, and then multiplying this total by the deviation. The summation of all such products results in Σab, which is measured in class intervals. Table LVI illustrates the derivation of Σab for Table LV.

TABLE LVI

Deviations A	Frequencies B +	Frequencies − C	Total D	Product + − D × A
1	47	9	+38	38
2	26	5	+21	42
3	4	0	+ 4	12
4	13	1	+12	48
6	16	0	+16	96
9	1	0	+ 1	9
12	1	0	+ 1	12
Totals	108	15	....	257
	15			
	127			
	$N = 250$		$\Sigma ab =$	25.7

As a check upon the inclusion of all frequencies in the computation, the plus (Column B = 108) and minus (Column C = 15) frequencies included in the above table may be added together. To this figure should be added the frequencies occurring within the zero zone (within the broken lines). In Table LV, 127

frequencies must be added, thus making the total N equal to the total frequency of the table.

According to the formula developed for the interval method, ab had to be derived from the *actual means*, whereas the sum 257 was obtained from assumed means. Therefore this total may not be correct unless the assumed means prove to be actual means. By referring to the computation on page 204 it is seen that such is not the case, the assumed mean group 66 to 67.9 inches is -0.22 *interval* too high. This is the correction K for the X series. The assumed mean weight group 140 to 149 pounds also has been assumed -0.048 interval too high, which is the correction K' for the Y series. Therefore, every product ab entering into the sum Σab which has been derived from the assumed center of gravity is incorrect to the extent of the product of the errors KK'. Since there are N such products the correction for Σab must be NKK', which must be subtracted from Σab when K and K' are of like sign, or added when unlike in sign. The numerator for the formula thus becomes $\Sigma ab - NKK'$, the sign of the product KK' indicating whether or not the minus sign before the N becomes plus.

The corrected formula then follows.

$$r = \frac{\Sigma ab - NKK'}{N\sigma_x\sigma_y}.$$

To avoid needless multiplication it may be simplified to

$$r = \left(\frac{\Sigma ab}{N} - KK'\right)\frac{1}{\sigma_x\sigma_y}, \tag{3}$$

in which Σab = sum of the products of deviations in intervals from assumed means.

N = number of paired items.

σ_x = standard deviation, in class intervals, of X series.

σ_y = standard deviation, in class intervals, of Y series.

K = correction, in intervals, for assumed mean for X series.

K' = correction, in intervals, for assumed mean of Y series.

By substitution for Table LV

$$r = \left[\frac{257}{250} - (-0.22 \times -0.048)\right]\frac{1}{1.276 \times 1.205} = 0.67.$$

The coefficient by this method is identical with that derived by the long method.

This is the easiest method for computing correlation in frequency series. The explanation may appear to make the method long and intricate, whereas in practice it is short and less liable to error. In its use, however, it should be remembered that all measurements are in *class intervals.*

108. Formulae.—The three formulae used in the study of correlation in this part are:

$$r = \frac{\Sigma_{xy}}{\Sigma_x^2} \tag{1}$$

All measurements are made in terms of the respective standard deviations. Used largely in time series and single frequency tables.

$$r = \frac{\Sigma_{xy}}{N\sigma_x\sigma_y} \tag{2}$$

The Pearsonian coefficient, the measurements are made from the means. Used preferably in single frequency tables.

$$r = \frac{\Sigma_{ab} - NKK'}{N\sigma_x\sigma_y} \tag{3}$$

All measurements are made in class intervals. An easy method for correlation tables.

A fourth formula may be derived from Eq. (2) by substitution for $N\sigma_x\sigma_y$. Since σ_x equals $\sqrt{\frac{\Sigma_x^2}{N}}$ and σ_y equals $\sqrt{\frac{\Sigma_y^2}{N}}$, $N\sigma_x\sigma_y$ must equal $\sqrt{\Sigma_x^2\Sigma_y^2}$, so that

$$r = \frac{\Sigma_{xy}}{\sqrt{\Sigma_x^2\Sigma_y^2}} \tag{4}$$

This formula saves one step in computation when compared to Eq. (2).

109. The Derivation of Correlation Coefficient by Graphic Methods.—A correlation table may be represented graphically by bringing into juxtaposition the plots of the two related phenomena. In order to distinguish the respective series, that is the length with its respective breadth from the breadth with its respective length, the plots of length, or X series, are represented by x's and those of breadth, Y series, by o's. *The scales upon which these data are plotted should be in the ratio of their respective standard deviations.* If the scales are so adjusted, then

formula $r = \frac{\Sigma_{xy}}{\Sigma_{x^2}}$ may serve as the basis of computation. To illustrate this method and principle a single frequency table is chosen in which, by previous computation (page 201), the correlation was found to be perfect. The standard deviations for these two series, as shown on page 200, are 5.66 inches for the length or X series and 2.83 centimeters for the breadth or Y series. The scale representing breadth should, therefore, be twice the size of that for length. Such scales are shown in Chart 52*a*.

Chart Based upon Table LI

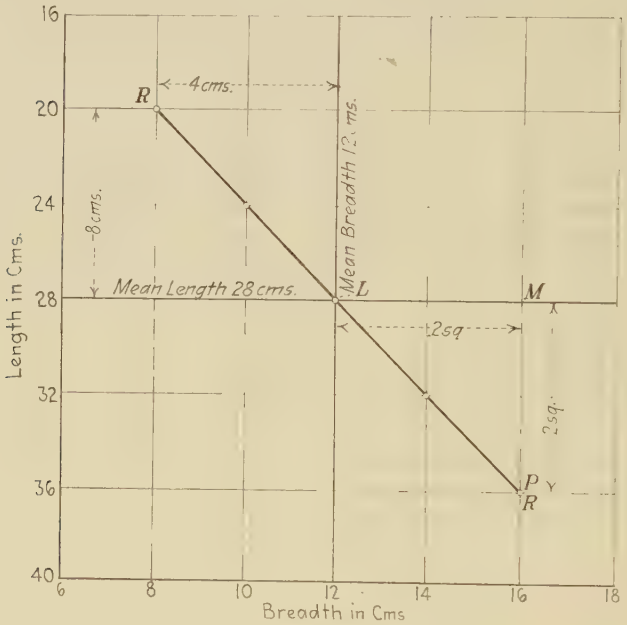


Chart 52*a*.

The data contained in Table LII are now plotted, an x being placed for every length with its related breadth, and o for every breadth with its related length. In a correlation table of more than single frequency these two plots are rarely identical. By observing the plot of one series only, length with its related breadth as represented by the x's, the associated value of y may be read from the value x, so that the product xy may be obtained for each of the paired items. The coefficient of correlation, however, is a value which expresses the relation which *on*

the average exists between two supposedly related series. The x's when plotted may scatter and not be on a straight line. This will occur if correlation is not perfect. Since correlation is an average relationship, a straight line, if the series is linear, must be drawn through the x's, so that the sum of the squares of their deviations from this line is at a minimum. This is the line of least squares for the X series. Such a line may be drawn by inspection or by mathematical precusion. The line so drawn represents at any point on it the average relationship of an x with its related y measured in terms of its standard deviations. Therefore any point on the line may be said to be representative of the relationship of the two series, so that the formula $r = \frac{\Sigma_{xy}}{\Sigma_{x^2}}$ may be written for graphic methods as $r = \frac{xy}{x^2}$, which by simplifying becomes $r = \frac{x}{y}$. This represents the tangent of the angle of the line from the center of gravity.

When the scales of the graph are in the ratio of the respective standard deviations, correlation may be measured by the value of the tangent of the angle of the line of best fit down through the plotted data.

In Chart 52*a* the x's lie on a straight line and afford no difficulty in drawing in the line RR which passes through the center of gravity a L. The tangent of the angle $MLP = \frac{PM}{LM} = \frac{2}{2}$ squares. The coefficient therefore is 1, indicating perfect correlation as derived in Table LII. *From this it appears that, when correlation is perfect and scales drawn in the ratio of the standard deviations, a line at* 45 *degrees to the mean lines indicates such a coefficient.*

By applying this method to Table LIII (page 203) it becomes apparent that, before any points may be plotted, a representative point has to be derived for each column and row of the table. In this problem let the attention be given to the X or height series. Since each plot has to represent both height and weight, a plot for height 60 to 61.9 inches must also indicate the related weight of the twelve persons found within this height group. The only representative point for weight would be the mean weight of the twelve persons. This is found to be 122.5 pounds. A plot placed at the intersection of the abscissa 61 and the

ordinate 122.5 would, therefore, represent graphically the class 60 to 61.9 inches with its related mean weight. By continuing the process for all classes in the X series points are plotted to which may be fitted the line of best fit. To obtain points for the Y series with its related x's the mean of the heights for the respective classes are computed. Both these computations are shown in Table LVII. It should be evident by this time that there are two lines for every series; X with its related y's and Y with its related x's.

TABLE LVII.—COMPUTATION OF THE MEANS OF ROWS FOR X SERIES AND MEANS FOR COLUMNS FOR Y SERIES OF TABLE LIII

X series		Y series	
Actual height	Mean weight	Actual weight	Mean height
60–61.9	122.5	100–109	61
62–63.9	133	110–119	63
64–65.9	143.9	120–129	63.2
66–67.9	143.5	130–139	65.1
68–69.9	151.3	140–149	66.8
70–71.9	160.8	150–159	68.2
		160–169	68.5
		170–179	70

Table LVII gives the points to be plotted on the graph which has its scales in the ratio of the respective standard deviations of the two series. The standard deviation of the X series was found to be (page 204) 2.552 inches and for the Y series 12.05 pounds. Thus the ratio of the scales should be 4.7 to 1, as illustrated in Chart 53.

The X series are represented by x's and the Y series by o's In the X series the points scatter and do not fall on a straight line, but the line of best fit ZZ is so placed as to reduce the sum of the squares of the deviations of the plots to a minimum. This line represents the degree of association between x and y for the X series. The tangent of the angle OLM of this line to the vertical equals $\frac{LM}{OM}$ equals $28\frac{1}{2}$ fiftieths of an inch, which gives a result of 0.67, which is identical with the computed correlation coefficient as on page 210.

A similar result is obtained for the Y series. The line of least squares, RR, makes a tangent of XOP or $\frac{PX}{OX} = \frac{27}{40.5}$, or a degree of association of 0.67.

The graphic method of the derivation of the correlation coefficient is more laborious than actual computation by formula and requires a knowledge of the line of least squares. It was developed here in order to introduce the subject of the lines of regression (RR and ZZ), since from these lines important values may be obtained. The line of perfect correlation is typified in Chart 53 by the broken line CC.

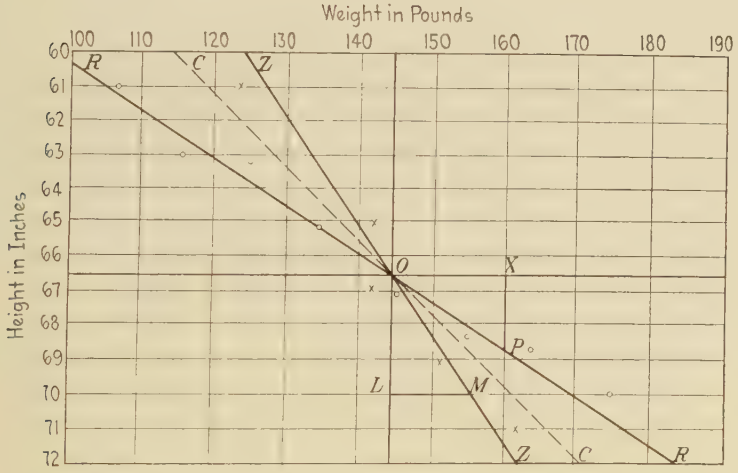


CHART 53.

110. Limitations of the Coefficient of Correlation.—From the discussion of the various ways of computing the coefficient of correlation it must be evident that correlation is a value which expresses the relation which *on the average* exists between two supposedly related series. It tells how the data disperse about the lines of best fit, as shown by the graphic method. If correlation is perfect, the data concentrate on a straight line, which indicates perfect synchronization of the fluctuations of the two series. If the relationship is less than 1, the lines of regression sweep between the means of the rows and columns. It is evident, therefore, that the series must be linear for the computation of the coefficient of correlation; that is, the data, when plotted,

must conform to straight lines and not be curvilinear. Series are not always linear, but in economic studies are usually so.

In a large number of cases, it is found either (1) that the mean of arrays lie very approximately around straight lines, or (2) that they lie so regularly (possibly owing to paucity of observations) that the real nature of the curve is not clearly indicated, and a straight line will do almost as well as any more elaborate curve. In such cases— and they are relatively more frequent than might be supposed—the fitting of straight lines to the mean of arrays determines all the most important characteristics of the distribution.[1]

However, for curvilinear series the computation by the straight line fit as typified by the correlation formula is, naturally, incorrect, and requires the use of the correlation ratio. The error involved by applying the straight-line fit to such series is shown in the "Primer of Statistics," by W. P. and E. M. Elderton, page 68. *Therefore, the coefficient of correlation will not measure association in series which are curvilinear.* Inasmuch as few series are markedly curvilinear the method of computation will not be given in this text.

A coefficient is merely a mathematical device and must not be used without discrimination, for

even a high degree of correlation may represent mere coincidence of phenomena. In a community in which the birth rate is declining many other tendencies will be in evidence which cannot be presumed to have any appreciable effect on the birth rate. A considerable degree of correlation might be found between the decline in the birth rate and the extension of railroad mileage, in a community over a long period of time, or between the birth rate and the increase in the number of letters sent through the post office. These correlations would be accidental and would not obtain in the annual fluctuations of birth, mileage, and letters. It is necessary to be constantly guarding against such accidental correlations in statistical work, since the statistical process, *being purely mathematical, does not discriminate between a causal and an accidental relationship.*[2]

Correlation gives an abstract value, and if the coefficient is less than 1, it indicates that the items scatter about the lines of best fit; that is, there is variation within one series. The coefficient is an average of these variations, but it may be important to be able to forecast relative changes in the two series;

[1] Yule, G. U., "An Introduction to the Theory of Statistics," p. 169.
[2] Bailey, W. and Cummings, J., "Statistics," p. 135.

that is, given a weight of 159 pounds it is required to know what the height would tend to be. The coefficient of correlation cannot answer this unless it is perfect. A new computation has to be made to indicate the ratio of variation.

The coefficient may arise through paucity of data or through obtaining poor samples of the phenomena. To illustrate, R. H. Hooker, in referring to the tossing of a coin, states:

> If we toss for a thousand times we shall almost certainly not get exactly 500 heads and 500 tails: we should get somewhere near those figures, and we should be very much surprised if we got 600 heads and only 400 tails. So great a difference as a proportion of 3:2 is very unlikely to occur, and if it did occur we should suspect very strongly that the tosser was able to exercise some (perhaps small) influence over the coin. Yet if we toss only five times, the nearest approach to equality that we can possibly get is 3:2, and if we toss ten times we are almost as likely to get 6:4 as 5:5. Thus, a distinct inequality in the case of a small number is of no significance, whereas it is important in the case of a large number. Similarly, a distinct correlation coefficient is no evidence of connection if the number of observations be small, but the same figure is if they be numerous. With a given coefficient, and a knowledge of the number of observations on which it is based, the probability of its being a true connection can always be calculated, and the "probable error" has been found to be the most useful guide. In practice it has been found to be *absolutely* safe to say that there is certainly a connection between two phenomena if the correlation coefficient be at least six times its probable error. This represents odds of many thousands to one in favor of the connection being real . . .
>
> Thus, whatever the correlation coefficient is (if not unity) there is a remote chance . . . and it may be millions to one . . . that the connection is accidental. But it is always strictly true that the coefficient measures the resemblance that has actually held for the observations included in the calculations. It is only when we want to know whether the connection is causal, and, as a corollary, whether it will still hold in the future, that we have to consider questions of probability, and it may perhaps not be amiss to remark that when a writer points out, as frequently occurs, that "inspection of the diagram shows ups and downs correspond fairly [Hooker is speaking of time series] and therefore the two sets of phenomena are interdependent" such a writer has frequently quite small odds in his favor.[1]

This quotation emphasizes the purely mathematical character of the coefficient and how it should be used with considerable care.

[1] HOOKER, R. H., "An Elementary Explanation of Correlation," *Quart. Jour. Roy. Meterolog. Soc.*, pp. 285–6, October, 1918.

In it reference is made to the "probable error" and the necessity for its computation.

111. *The Probable Error.*—The probable error, as already indicated, gives the limits within which the coefficient of correlation may be said to fluctuate owing to the possibility of defective sampling. The formula for such is

$$\text{Probable error for } r = 0.6745 \times \frac{1 - r^2}{\sqrt{N}},$$

in which 0.6745 is a constant.

r = coefficient of correlation.
N = number of paired items.

By substitution for values in Table LVI

$$r = 0.6745 \times \frac{1 - 0.67^2}{\sqrt{250}} = 0.02.$$

The probable error for the table is found to be 0.02 for the coefficient of 0.67. Since the coefficient may be too small or too large by this amount, owing to defective sampling, the coefficient should be written as

Coefficient of correlation for Table LVI = 0.67 ± 0.02, which indicates that the measure of association may fluctuate between 0.69 or 0.65. From this it is apparent that no correlation coefficient should be given as one arbitrary value, such as 0.67 (unless correlation is perfect), but rather given at the calculated value with an expression of the probable error, as

$$\text{Correlation} = 0.67 \text{ probable error} = 0.02$$

From Hooker's[1] quotation it is also learned that the coefficient of correlation should be a value which exceeds six times the probable error, if it is to be considered significant. To this may be added another arbitrary rule— the coefficient to be significant should be at least 0.5.

112. The Coefficient and Lines of Regression.—In related series it is important to determine "how great a divergence of the one variable from its average value is associated with a unit divergence of the other, and to obtain some idea as to the closeness with which this relation is usually fulfilled."[2] The closeness with

[1] Hooker, R. H., "An Elementary Explanation of Correlation," Quart. Jour. *Roy. Meteorolog. Soc.*, pp. 285–6. October, 1918.

[2] Yule, G. U., "An Introduction to the Theory of Statistics," pp. 167–168.

which this relation is usually fulfilled is the measure of correlation, while the measure of variation is called the coefficient of regression. The two measures are therefore complementary to one another in the analysis of the association of series, although some authors give first place to regression.

The regression coefficient is statistically the more fundamental and in all actual problems involving the estimate of one variable, knowing a second, the regression coefficient and not the correlation coefficient is the essential measure.[1]

Correlation and regression coefficients are not identical, as may be seen in Table LVIII and LIX.

TABLE LVIII

X	*Y*
Length in centimeters	Breadth in centimeters
20	8
24	10
28	12
32	14
36	16

TABLE LIX

X	*Y*
Length in centimeters	Breadth in centimeters
10	2
12	2.5
14	3
16	3.5
18	4

In these tables the correlation is perfect, the respective coefficients being 1. Nevertheless, the ratios of variation are not

[1] KELLEY, TRUMAN L., "Statistical Method," p. 181.

identical. A unit increase of length in Table LVIII has a corresponding increase of breadth of 0.5 inch, while in Table LIX, the same increment increase in length has a corresponding change in breadth of only 0.25 inch. The ratios of length to breadth are 2 to 1 and 4 to 1 respectively. Ratios may, therefore, be any value, but if in any correlated series the ratio of X to Y is greater than unity, the ratio of Y to X must be less than 1. Therefore, the coefficient of correlation is not the same thing as the "ratio of variation" or, in more technical terms, the coefficient of regression. If the coefficient of correlation for any table is less than 1 it in itself affords no basis for measuring the ratio variation.

The lines of regression are familiar, having been given in Chart 53. The lines ZZ and RR, however, were drawn by the method of least squares, but the result is identical with that of the regression method.

Since there are at least two series to every correlation study there must be two regression equations which will estimate: (1) the change in the Y series for a unit change in the X series, and (2) the change in the X series for a unit change in Y. The two equations will, naturally, be a function of the coefficient of correlation, because it expresses to what degree, in general, there is association. It has been stated that when correlation is *perfect* it is possible to predict the ratio of variation in the two series; that is, given a 4-inch increase in length (Table LVIII), an increase in breadth of 2 inches can be predicted.

If correlation is perfect and the standard deviations of the respective series known, it should be apparent that the change (from the respective means) in Y for a unit change in series X will be in the ratio of their respective standard deviations. Perfect correlation implies perfect synchronization of change in the related series. It does not, however, imply that a unit change in X from its mean is followed by the same unit change in Y. It does indicate, however, *a constant ratio change*. Two series which are perfectly correlated might have standard deviations in the ratio of 2 to 1, the X series having twice the measure of dispersion of the Y. Naturally, twice the dispersion implies that the spread from the mean of X series is twice that of Y. Therefore a unit of change in X from its mean will correspond to 0.5 unit change in Y. In other words, the change will be according to the ratio of the standard deviations. In Table LVIII,

it will be found that the ratio of variation is according to the ratio of the standard deviations of the related series. The standard deviation for length, or X series, is 5.86, while that for breadth, or Y, series, is 2.83 centimeters. A deviation of 1 from the mean of the X series will yield a deviation of $\frac{2.83}{5.66}$, or 0.5 in Y.

This result is obtained because correlation is perfect. Naturally, if association were not perfect the ratio of the standard deviations would have to be corrected according to the coefficient of correlation since correlation is a measure of the average degree of association. The formula for the derivation of the change in Y from its mean for a given change in X from its mean becomes, therefore

$$y = r \times \frac{\sigma_y}{\sigma_x} \times x, \tag{1}$$

in which y = the change from the mean of the Y series for a given change in X.
r = coefficient of correlation.
σ_y = standard deviation of Y series.
σ_x = standard deviation of X series.
x = given change from mean of X series.

Substituting the values of Table LVIII

$$y = 1 \times \frac{2.83}{5.86} \times 1 = 0.5.$$

The result 0.5 means that, in the X series, for every unit of change from its mean there is, on the average, in the Y series, 0.5 unit change from its mean.

If it were required to find the change in X for a unit change in Y the formula would become

$$x = r\frac{\sigma_x}{\sigma} \times y \tag{2}$$

Substituting values for Table LVIII

$$x = 1 \times \frac{5.86}{2.83} \times 1 = 2,$$

which indicates for every unit change from the mean of the Y series there is twice the change in the X series.

The two equations are called regression equations and from the formulae and reasoning it is seen that their values are depend-

ent upon both the ratio of the respective standard deviations and the coefficient of correlation. These equations when expressed as coefficient are usually written

$$b_1 = r\frac{\sigma_x}{\sigma_y} \tag{3}$$

$$b_2 = r\frac{\sigma_y}{\sigma_x} \tag{4}$$

b_1 indicates the change in X series for a given change in Y. b_2 indicates the change from the mean of Y for a given change in X.

Application of Eqs. (3) and (4) to Table LV, which has a coefficient of 0.67, standard deviation, for X or height series of 2.552 inches, while that for Y or weight is 12.05 pounds, that is

$$b_1 = 0.67 \times \frac{2.552}{12.05} \qquad b_2 = 0.67 \times \frac{12.05}{2.552}.$$

Therefore

$$x = 0.14y \qquad y = 3.1x$$

These equations mean that the deviation of x for the X series will be 0.14 that of Y change, or change of 1 pound in weight corresponds to 0.14 change in height. The second equation indicates the deviation y from the mean of the Y series to be 3.1 that of x for the X series, or a change of 1 inch in height corresponds to a 3.1-pound change in weight.

In order to answer the question, what height would correspond on the average to a weight of 150 pounds? It is necessary to express x and y not as deviations from the means of the respective series but in absolute values. Since x and y represent actual deviations from the actual means of the X and Y series respectively, they must equal the difference between any arbitrary assumed value for the series and their respective means. Therefore,

$$x = X - \text{Mean for } X \text{ series}$$
$$y = Y - \text{Mean for } Y \text{ series}$$

X and Y being arbitrary values for the respective series. The derivation of values can perhaps best be indicated by application to Table LV.

Application to Table LV.—Given
Mean height 66.5 inches (page 204) X series
Mean weight 144.5 pounds (page 204) Y series

Therefore,

$$x = X - 66.5 \qquad y = Y - 144.5 \qquad (a)$$

By substituting values for x and y from regression equations $x = 0.14y$ and $y = 3.1x$ (page 222)

$$0.14y = X - 66.5 \qquad 3.1x = Y - 144.5 \qquad (b)$$

but $y = Y - 144.5$ and $x = X - 66.5$ (row a), so that

$$0.14(Y - 144.5) = X - 66.5 \quad 3.1(X - 66.5) = Y - 144.5 \quad (c)$$

Therefore

$$X = 46.5 + 0.14y \qquad (d)$$
$$Y = 3.1x - 61.65 \qquad (e)$$

It is from Eqs. (d) and (e) that absolute values may be found. When the weight Y is 150 the height will be, by substitution of value Y in formula (d), 67.5 inches, for when Y is 150

$$X = 46.5 + 0.14(150)$$
$$= 67.5 \text{ inches}$$

The result, 67.5 inches, indicates that, on the average, such a height corresponds with a weight of 150 pounds when computed from such data as in Table LVI.

From the regression equations it becomes possible to find the ratio of change between the related variables and by the substitution of the actual means in the equations it becomes possible to determine from known value of X the unknown value for its associated Y, and *vice versa*.

113. Graphic Presentation of Lines of Regression.—The graphic presentation of the lines of regression is shown in Chart 54. This chart is identical with that on page 215, the only difference being the method of derivation. In this chart the regression equations form the basis of drawing in the lines of regression. Two arbitrary values were assumed for both X and Y which gave two points on the chart through which were drawn the lines of regression RR and ZZ. The derivation of the points is indicated below the chart, the points being represented on the graph by four dots. Naturally, the tangent of the angles of these lines to their respective means will be 0.67 the value of the coefficient of correlation, since they were derived from the coefficient.

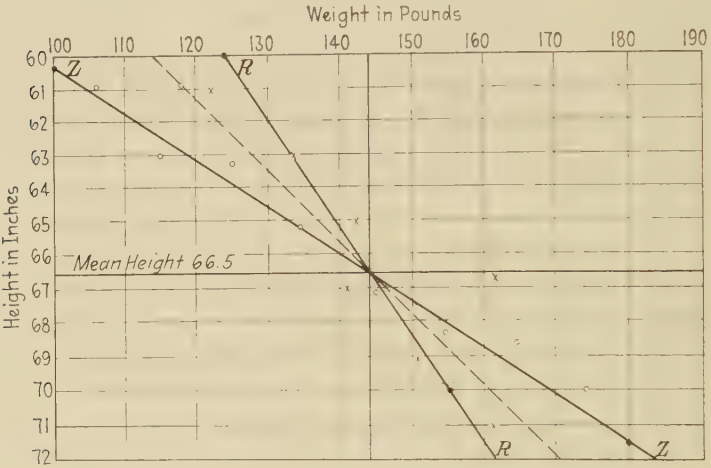
114. Summary.—Correlation is an important measure in statistical studies, but it should be used with extreme care, since

it is purely a mathematical device and does not allow for accidental relationships, or defects through the paucity of data. Its measures in order to be significant should be at least six times its probable error and have a minimum value of 0.5.

The ratio of variation requires the use of regression equations since the correlation coefficient cannot give this measure.

The coefficient of correlation does not suffice to characterize a series the data of which are curvilinear.

LINES OF REGRESSION FOR TABLE LV



Line *ZZ*

$X = 46.5 + 1.14y$

When $y = 100$

$X = 46.5 + 14 = 60.5$

When $y = 180$

$X = 46.5 + 25.2 = 71.7$

Line *ZZ*

$Y = 3.1x - 61.65$

When $X = 60$

$Y = 186 - 61.65 = 124.65$

When $X = 70$

$Y = 217 - 64.65 = 155.35$

CHART 54.

References

BAILEY, W. and CUMMINGS, J.: "Statistics."

BOWLEY, A. L.: "Elements of Statistics," pp. 316–327.

BRAVAIS, A.: "Analyse mathématique sur les probabilités des erreuor de situation d'un point," *Acad. Sci. Mem. (presentés par divers savante)*, p. 255, 1846.

BROWN, W.: "The Essentials of Mental Measurement."

DAVIES, G. R.: "Introduction to Economic Statistics," pp. 131–144.

ELDERTON, W. P. and E. M.: "Primer of Statistics," chaps. V and VI.

GALTON, SIR FRANCIS: "Family Likeness in Stature," *Proc. Roy. Soc.*, vol. 40, pp. 63–73, 1886.

HOOKER, R. H.: "An Elementary Explanation of Correlation," *Quar. Jour. Roy. Meteorolog. Soc.*, vol. 34, pp. 277–291, 1908.
KELLEY, TRUMAN L.: "Statistical Method," chap. IX and X.
KING, W. I.: "Elements of Statistical Method," pp. 197–202.
PEARSON, KARL: "Regression, Heredity, and Panmixia," *Phil. Trans. Roy. Soc. (London)*, Series A, vol. 187, pp. 253–318, 1896.
PERSONS, WARREN M.: "Construction of a Business Barometer Based upon Annual Data," *Amer. Econ. Rev.*, December, 1916, pp. 767–769.
SECRIST, H.: "An Introduction to Statistical Methods."
——: "Readings and Problems in Statistical Methods," pp. 396–416.
WEST, CARL: "Introduction to Mathematical Statistics," chap. VII–IX.
WHIPPLE, G. G.: "Vital Statistics," chap. XIII.
YULE, G. U.: "An Introduction to the Theory of Statistics," chap. XVIII.
——: "The Applications of the Method of Correlation to Social and Economic Statistics," *Jour. Roy. Stat. Soc.*, pp. 721, December, 1909.

Problems

1. What is correlation?

2. Why does the coefficient of correlation never exceed a value of 1?

3. What is meant by the terms "positive" and "negative" correlation?

4. How may the approximate measure of correlation be derived graphically?

5. Record opposite each age the respective straight life insurance premiums for the ages forty to sixty-five inclusive.

6. From the simple correlation table in problem 5 construct a graph so that the approximate degree of correlation is obtained.

7. Check the result of problem 6 by computing the coefficient of correlation by the standard deviation method, also by the Pearsonian formula.

8. Determine by the interval method the coefficients of correlation for the two correlation tables on the vote in Cuyahoga County, Ohio, as found in Chap. XVII.

9. From the gymnasium records obtain the weight with its corresponding height for each of 100 students.

10. Tabulate the results of problem 9 into a correlation table which has intervals of 5 and 2 for weight and height respectively.

11. Compute, by the interval method, the coefficient of correlation.

12. *a.* Is the coefficient of problem 11 significant? Why?

b. Would it have been significant, even though high in value, if only ten samples had been taken and not 100? Explain.

13. How may the coefficient be corrected for errors in sampling?

14. *a.* Of what significance is the probable error?

b. Compute it for all problems given above.

c. List the corrected coefficients of correlation.

15. What are the limitations of the coefficient of correlation?

16. What are the important attributes of the regression equations?

17. *a.* How could you, in a correlation study, determine the value of Y for a known value of X?

b. Apply the answer to problem 9.

18. Represent graphically the lines of regression of problem 9.

PART III

STATISTICAL METHODS FOR TIME SERIES

CHAPTER XIX

TIME SERIES, THEIR PROBLEMS

Much time and energy are now being given to the study of those phenomena which vary over a period of time. As already mentioned elsewhere, statistical control is the slogan of practically every large business concern. Statistical control does not mean merely the knowledge of output, but rather the knowledge of present and future market conditions, so that output may be regulated accordingly. The object is to be able to forecast business conditions so that the ever-recurring periods of maladjustment between production and consumption may disorganize business less severely. Production, it is said, is carried on in anticipation of demand. The problem, therefore, is to measure, if possible, the future demand. Demand, it is true, follows no rigid law, but rather the whims of society and, to a greater or lesser extent, those of the advertising managers. In spite of the difficulties, however, a study of past and present wants with their interrelationships may afford a good basis for estimating future needs. If such is the case, time series must be analyzed.

The regulation of industry, both seasonal and cyclical, requires two instruments of measurement; one a thermometer to measure the rise and fall of the industrial temperature from normal, and the other, a barometer, by means of which weather conditions of business may be forecasted and the elevation reached in the business cycle measured.

The thermometer of business often takes the form of index numbers, which are usually averages of certain phenomena, at a given date, expressed as percentages of the same phenomena measured at a base period, *e.g.*,

$$\frac{\textit{Mean of Wholesale prices,}\ \text{September, 1924}}{\text{Mean of wholesale prices, 1913}} = \text{Index number.}$$

An index is, as is the thermometer in the hands of a physician, merely a guide. It alone cannot indicate all the symptoms,—it must be supplemented by further tests.

The phenomena of the business cycle are familiar to all. Types of business fluctuate in various degrees within any one

year, thus giving rise to seasonal fluctuations. Likewise, there is a growth element over a period of time. Pig-iron production increases year after year because of the growth of business, which, in turn, is dependent somewhat upon the growth of population. These gradual yearly accretions indicate positive secular trend. Industries may also decline with time, indicating a negative trend. Secular trend, therefore, is the constant accretion or diminution in a series which takes place over a long period of time. The periods of prosperity and depression are too well known, and these oscillations are called "cyclical fluctuations." A time series, then, *may* have seasonal, secular, and cyclical variations. Fortuitous variations may also be present.

The business barometer, by eliminating secular and seasonal variations and leaving only the cyclic, endeavors, by means of computing the coefficient of correlation between series having different degrees of lag, to derive a forecaster. A barometer is an instrument which not only forecasts weather conditions but also affords a basis for the determination of land elevations. Business needs a similar measure to indicate what may happen in the future, and also, at the same time, to give bearings of the position now occupied in the business cycle. This twofold need gives rise to the use of index numbers in the construction and use of a business barometer. The index numbers indicate the elevation reached.

The readings given by forecasters should, if properly constructed, warn business when to prepare for changed conditions. There are agencies which make a specialty of forecasting services. The development of forecasting has been phenomenal, but as yet it is in its infancy. The general public realize that weather barometers are not always trustworthy, as weather conditions may change very rapidly. So, too, with business forecasters. Other conditions are never equal. Therefore, predictions will sometimes be contrary to actual business conditions.

These two types of measures of time series form the basis of the discussion of Part III. The measures of position, dispersion, and skewness are not taken up in detail because their derivation should be apparent from the study of these measures in Part II. If such measures are required for time series, the data are rearranged, chronology losing its significance in the stub of the table. Such measures are easy of computation, since the frequency of a time series is one, there is but one value for the given year.

CHAPTER XX

INDEX NUMBERS

115. Definition.—An index number is the number which characterizes the degree of change, either over a period of time or between series at the same time, of a group of items when measured from some predetermined standard or base, the base usually having a designated value of 100. Reverting to analogies, it is the reading of the temperatures from normal, which for the physician is 98.6, but for the statistician 100 (except where aggregate values are used). Any value of a group, irrespective of its size, units of measurement, or time, if it is to be considered the base, may be given a value of 100. Index numbers are not relative numbers, since the latter have a definite relationship to a *specific* datum, whereas index numbers express a *general relationship* to all data contained within the index group. An index is a group figure and may not, necessarily, be represented by any specific data within the group, since it is usually an average of relatives. Professor Fisher in speaking of price indices defines them as follows:

> An index number of prices, then, shows the *average percentage change* of prices from one point of time to another. The percentage change in the price of a *single* commodity from one time to another, is, of course, found by dividing its price at the second time by its price at the first time. The ratio between these two prices is called the *price relative* of that one particular commodity in relation to those two particular times. *An index number* of the prices of a *number* of commodities is an *average* of their price relatives.[1]

116. Brief Early History.—The business man has access to various indices, such as those of physical production, volume of trade, wholesale and retail prices, real and money wages, cost of living, building costs, stocks and bonds, and many others. Many of these indices are of comparatively recent origin. Therefore,

[1] Fisher, Irving, "The Making of Index Numbers," p. 3.

it is often assumed that the device used in measuring group changes is a direct product of the Industrial Revolution. It is true that it is a new tool for the statistician, its general use dating from about 1900 with rapid development during and after the Great War. But it is really a tool refashioned from a rough-hewn device dating back to the eighteenth century.

The early history of index numbers is closely associated with price movements. The discovery of the New World in 1492 led to a phenomenal change in prices, when the precious metals poured into the coffers of Spain and from there were circulated throughout the rest of Europe. Price phenomena began to attract the attention of statesmen and students. Dutot, a Frenchman, in 1738, by summating the prices of a list of goods if purchased in the time of Louis XII (1498 to 1515) compared them with the prices of the same list of articles if they had been purchased in the time of Louis XIV (1661 to 1715). By comparing these simple aggregate figures an idea of the purchasing power of money was obtained. The first study to derive an index through relative numbers is attributed to Carli, who, in 1764, endeavored to show the change in the purchasing power of money from 1500 to 1750, as indicated in the prices of grain, wine, and oil. The method used in deriving the index was that of the simple arithmetic mean of the price relatives. Price relatives of the three commodities were obtained by dividing the cost of each in 1750 by its cost in 1500, the base year. The relatives so obtained were then summated and the mean of the sum derived, which gave the index.

The idea that the data comprising an index could have relative degree of importance was first worked out in a series by Arthur Young, who in 1812 arbitrarily assumed that the degrees of importance of certain goods were as follows: wheat, 5; day labor 5; provisions, 4; barley, 2; oats, 2; wood, coal, and iron, 1 each. In other words, he gave weights to the articles, as indicated by these numbers. The respective quotations were multiplied by these numbers. The resulting products, when summated and divided by the sum of the weights, gave the weighted mean index. His system of weighting was, indeed, crude, but it emphasized the necessity of judging the relative importance of the constituent parts of an index.

Another contribution was made to the development of index numbers when it was recognized that the simple arithmetic

mean exaggerates a rise and minimizes a fall in the data if the recent is compared with the more remote. Let it be supposed that, in 1924, the price of a commodity X has doubled since 1913, while that of another commodity Y has fallen to one-half during the same period. Their relative numbers in 1924 to the 1913 base will then be 200 for X and 50 for Y the arithmetic mean of which equals $\frac{250}{2} = 125$. Such an index indicates the prices to have increased, in general, 25 per cent during the last ten years. If these two articles are of equal importance in the family budget, it is self-evident that the cost of living has not risen but is still at 100, the same as in 1913. W. S. Jevons in 1865 made a study of prices in England and advocated the use of the geometric mean, a mean which is derived by multiplying the relatives together and extracting the *n*th root (according to the number of relatives) of the product. This method would avoid the exaggeration of increase in prices as shown above, for, by application,

$$\text{Geometric mean} = \sqrt{200 \times 50} = 100.$$

This mean, however, has never come into general favor, but by the efforts of Jevons considerable interest was created in the use of index numbers.

Subsequent history records great achievements in the use and development of index numbers. It is not the intention, however, to give a history of index numbers, but only to emphasize, once more, the fact that statistics originated in a period which antedates the industrial era. Mention was made of W. S. Jevons because he has been called the "father of index numbers."

117. Summary of Problems Involved in the Construction of Index Numbers.—The study of index numbers was reserved for Part III not because index numbers always require comparison over a period of time, for such is not the case, since it is possible to compare them at any *given* time; but because they generally measure group changes over time.

Since index numbers apply to groups of data, problems naturally arise pertaining to the constituent parts of the group, and also in connection with the methods of analysis to be used. These problems may be summarized as follows:

- Component parts
  - *a.* Types of data to be included.
  - *b.* The selection of samples.
  - *c.* The number of samples.
  - *d.* Relative importance of the samples.
- Analysis
  - *a.* Base for comparison
    - Fixed.
    - Chain.
  - *b.* Methods for comparison
    - Simple and weighted
      - Arithmetic mean.
      - Geometric.
      - Harmonic.
      - Median.
      - Mode.
      - Aggregative.

These problems will be discussed in the order given.

118. Types of Data to Be Included in Index.—As in any other statistical problem, the type of data to be used in any study is determined by the purpose of the study, purpose determining both the contents and analysis.

Index numbers may be divided into two types: (1) general-purpose indices, and (2) special-purpose indices. General-purpose indices are those which tend to show price levels (if a price index) in general, not those of a particular group of commodities, such as farm products, which is a special purpose type. Naturally, the inclusion and exclusion of data for the index depends upon the purpose—a general-purpose table would have a much wider range of commodities than a specific index.

The first rule in the construction of a group index is to have *the purpose definitely in mind, together with rigidity in the definition of the unit of measurement.* In constructing a cost-of-living index the purpose clearly indicates that only those things should be included which enter into a family budget. Furthermore, the unit of measurement should be retail prices, the price paid by the consumer himself. The term "family" should be defined before collection of data is begun. Then follows the problem of deciding what commodities should be included under the following group headings. Food, Clothing, Rent, Heat and Light, Furnishings, and Miscellaneous. The Bureau of Labor included the following items in the Food group of the Cost of Living Index:

TABLE LX.—FOOD WEIGHTS—ANNUAL CONSUMPTION PER FAMILY[1]

Article	Unit	United States	North Atlantic	South Atlantic	North Central	South Central	Western
Sirloin steak	Pound	32	27	35	34	38	39
Round steak	Pound	32	27	35	34	38	39
Rib roast	Pound	31	30	24	32	24	39
Chuck roast	Pound	31	30	24	32	24	39
Plate beef	Pound	23	25	17	23	16	27
Pork chops	Pound	36	29	43	45	42	25
Bacon	Pound	17	13	20	18	17	19
Ham	Pound	22	26	43	14	19	10
Lamb	Pound	8	14	2	2	1	13
Hens	Pound	23	25	24	23	22	19
Salmon, canned	Pound	9	10	9	9	9	6
Milk, fresh	Quart	337	412	155	364	177	377
Milk, evaporated	Pound	77	95	73	48	85	92
Butter	Pound	66	75	56	53	60	89
Oleomargarine	Pound	16	8	9	30	16	8
Nut margarine	Pound	6	4	5	11	3	2
Cheese	Pound	12	12	13	12	11	15
Lard	Pound	34	27	38	45	38	18
Crisco	Pound	9	6	10	5	22	16
Eggs, strictly fresh	Dozen	61	68	57	53	55	70
Bread	Pound	531	642	417	521	450	438
Flour	Pound	264	224	313	263	318	280
Corn meal	Pound	54	29	108	39	140	34
Rolled oats	Pound	41	45	31	39	38	45
Corn flakes	Pound	7	6	6	6	13	5
Cream of Wheat	Pound	7	7	2	6	3	12
Macaroni	Pound	23	25	15	20	29	27
Rice	Pound	35	32	55	26	56	28
Beans, navy	Pound	22	23	17	25	21	19
Potatoes	Pound	704	746	514	810	485	706
Onions	Pound	66	72	52	62	82	64
Cabbage	Pound	65	62	61	70	66	61
Beans, baked	Pound	7	8	10	6	5	4
Corn, canned	Pound	10	8	9	13	10	10
Peas, canned	Pound	10	10	9	13	9	9
Tomatoes, canned	Pound	16	15	21	10	35	12
Sugar	Pound	147	140	145	154	133	161
Tea	Pound	8	13	6	5	3	6
Coffee	Pound	40	33	42	45	52	35
Prunes	Pound	11	14	9	11	8	10
Raisins	Pound	9	9	4	11	7	12
Bananas	Dozen[2]	11	11	8	11	13	9
Oranges	Dozen	7	6	9	6	9	8

[1] U. S. Bureau of Labor Statistics, *Bull.* 334.

[2] In cities where most of the sales on bananas are by the pound rather than by the dozen, the weightings as given in the above table have been multiplied by 3 and have then been applied to the prices on the pound.

By giving to each article a weighting equal to its relative importance in the consumption of the average family, the total expenditure for certain articles of food on a given date forms a proper basis of comparison with the expenditure for the same articles of food on any other date, and for the purpose of showing the movement in retail prices it is assumed that this relative importance remained the same through the whole period, 1913 to 1921. In this bulletin the average family expenditure in each city in the year 1913 has been taken as the base for that city, and index numbers have been computed for each month, January, 1913, to December, 1922, and for each year, 1914 to 1922. These index numbers show the trend in the retail cost of all foods combined in each individual city as compared with the average cost in that city in the year 1913.

An index of real wages would include a study of the cost of living on the one hand, and that of money wages on the other, but the term "wage" should be defined, since it might mean hourly, weekly, actual earnings, or potential wages.

If the purpose of an index is to construct a business barometer, only those goods exhibiting definite cyclical fluctuations should be included. Farm products as a group should be excluded, since they do not synchronize in any marked way with changes in business conditions. This fact must be self-evident from the mass of farm-relief legislation which is still (1924) advocated. A general-purpose index which has been wrongly used as a rough barometer has been that of wholesale prices as published by the Bureau of Labor Statistics. This is essentially an index measuring price levels and not general business conditions. As now published it contains "404 commodities or price series," the index being derived from subindices for the different groups of commodities as shown in Table LXI.

TABLE LXI.—INDEX NUMBERS OF WHOLESALE PRICES, BY GROUPS OF COMMODITIES

Group	1923		January, 1924
	January	December	
Farm products	143	145	144
Foods	141	147	143
Cloths and clothing	196	203	200
Fuel and lighting	218	162	169
Metals and metal products	133	142	142
Building materials	188	178	181
Chemicals and drugs	131	130	132
House-furnishing goods	184	176	176
Miscellaneous	124	116	117
All commodities	156	151	151

Monthly Labor Review, February, 1924.

Wesley C. Mitchell in writing of the Wholesale Index states:

Probably the most illuminating way of presenting an index number that aspires to cover the whole field of prices at wholesale would be to publish separate results for the groups that have characteristic differences of price fluctuations, and then to publish also a grand total including all the groups. The groups to be recognized and the distribution of the commodities among them are difficult matters to decide. But, as matters stand, the most significant arrangement seems to be (1) a division of all the commodities into raw and manufactured products; (2) the subdivision of raw commodities into farm crops and animal, forest, and mineral products; (3) the subdivision of the manufactured products according as they are bought mainly for personal consumption, mainly for business use, or largely for both purposes.[1]

The subindices of the wholesale price index are considered by some to indicate and forecast general business conditions. A rather unique study of wholesale prices was that made by Joseph Snider, who in analyzing wholesale prices from 1866 to 1891 stated that "one of the primary objects of this study was to ascertain which commodities showed similar price movements

[1] "Index Numbers of Wholesale Prices," U. S. Bureau of Labor Statistics, *Bull.* 284.

during the period."[1] The results of this study, while not conclusive, show the so-called economic groupings, as advocated by Wesley C. Mitchell, to be very dissimilar in their statistical properties. Therefore, the use of subindices or the general index of wholesale prices for forecasting is more or less vitiated. The differences in fluctuations of commodities contained within the same group is clearly indicated in Table LXII, the captions of which indicate the graphic variations.

TABLE LXII.—TWO-WAY CLASSIFICATION OF THE SERIES

Group	Cyclical A	Cyclical B	Cyclical C	Seasonal	Inflexible	Irregular, trend broken, A	Irregular, trend broken, B	Irregular, trend unbroken	Miscellaneous	Total
Raw materials:										
Farm crops	3	..	5	1	..	1	1	4	2	17
Animal products	1	1	1	3	1	..	1	1	2	11
Forest products	..	..	..	..	2	..	..	..	..	2
Mineral products	..	..	..	..	..	..	1	1	2	4
Manufactured products:										
Producers' goods	..	4	3	..	2	2	5	..	10	26
Consumers' goods	3	..	1	3	8	6	..	6	..	27

By permission of Joseph Snider.

The analysis of this table shows clearly that a general wholesale index is unreliable criterion of business changes.

The *Harvard Review of Economic Statistics* has recognized the wide diversity of price fluctuations in grouped commodities and has developed a "Ten Commodity Index" which forms part of their business barometer. The commodities selected are those which have marked cyclical fluctuations and are more sensitive to business conditions than those included in the larger wholesale index, which includes many but diverse commodities. The commodities included in this business index are: cottonseed oil, coke, pig iron, bar iron, mess pork, hides, print cloths, sheetings, and worsted yarns. The purpose of this index, therefore, determines the inclusion and exclusion of certain types of price data. Thus its contents differ radically from that of the wholesale index, even though it might include commodities which are found in the wholesale index.

[1] SNIDER, JOSEPH, "Wholesale Prices in the United States, 1866–1891," *Rev. Econ. Stat.*, April, 1924.

An index of building costs, such as the Aberthaw Index, suggests at once the inclusion of its costs of building a certain type of structure. The chief problem in such an index arises in defining the term "cost," since it might mean market, jobber's, contract, or retail. The index as constructed used none of these suggested definitions, but stated the cost to be the best price that the Aberthaw Company could get on its materials. The base used for comparison was the itemized cost in 1914 of buiding a certain type of seven-story, reinforced-concrete factory building. The itemized cost of building the same structure, at various periods of time, when divided by the base cost, results in the building index.

Naturally, in making an index of collections for a business house the problem would require the measurement of the relationship between total sales and outstanding accounts. Rigidity of the definition of the time criterion of the term outstanding would be necessary, likewise the rules for handling the discounts granted.

An index of the volume of trade differs from that of physical production, since the term "trade" includes only those things which are *actually exchanged,* whereas physical production includes all things produced. The total corn crop would be included in the physical index even though the greater part of it is never exchanged directly, but rather indirectly in the form of pork. On the other hand, an index of trade may include commodities imported, whereas that of physical volume would include only those goods produced within the boundaries of the district or country.

The Retail Price Index of the United States Bureau of Labor Statistics has for its object the publishing of index numbers of retail prices for cities, geographical divisions, and the entire United States. In such an immense study the problem of what data to include is an enormous one, since the final index is supposed to be representative of the entire country. The purpose decrees the collection of data for cities from which are derived the geographic index, which in turn forms the basis of the final composite index.

The foregoing illustrations have been given in order to emphasize the need for the careful selection of data to be included in any index. Even in general-purpose indices judicious selection of the types of data represented is imperative. The illustrations

also emphasize, on the other hand, that index numbers cannot be used indiscriminately. Their contents should always be known. To follow this rule implies that all index numbers give the source of their material and contents. Unfortunately, this is not a common practice. Progress, however, is being made in this direction.

To summarize in a little different way, it has been found that:[1]

1. While fluctuations in raw materials are found in their manufactured products, the changes in the latter tend to be less pronounced.

2. Farm crops have far wider fluctuations in price than forest, mineral, and animal products because farm products depend to a great extent upon climatic conditions.

3. The price of mineral products fluctuate more in conformity with business conditions than do forest or animal product prices.

4. Manufactured goods used by producers fluctuate more in price than consumers' goods, as retail prices are influenced more by custom.

5. The prices of the various commodities contained within any one of the above-mentioned groups vary a great deal, there being found different types of cyclical fluctuations, some having seasonal variations while other have "irregular" and "inflexible" trends.

Variations, therefore, are found within ostensibly the same data, and, since an index number is the measure of a composite group, the groups of data included should conform to the purpose of the index. Care in selection of data is extremely important in the construction of an index, and the selection depends largely upon the purpose of the index.

119. The Selection and Number of Samples to Use.—*a. The Types of Samples.*—The problem discussed in connection with the selection of the groups of data to be included within an index would be the only problem prior to analysis if

the price fluctuations (assuming it be a price study) of each commodity that is bought and sold were strictly independent of the price fluctuations of every other commodity, and, if (2) each commodity has just the same importance as an element in the general system of prices as every other commodity, then any series of price quotations collected at random would be a fair sample for determining the average changes in the

[1] Based on MITCHELL, WESLEY, C., "Index Numbers of Wholesale Prices," and SNYDER, JOSEPH, "Wholesale Prices in the United States."

wholesale prices of commodities in general. Of course, the larger the number of commodities included the more trustworthy would be the index number.[1]

Price phenomena, however, are a complex of interrelations, and commodities do vary in their relative importance, so that the problem of selection of data from the predetermined groups of data to be represented in the index is not a simple matter. The samples chosen should be representative of the entire group, so that the index may represent all the types of data which could be contained within the group, if it were possible to collect them. A good guide in determining the type of samples to use is that "each is as little correlated as possible with the other quotations included in the index, and (*b*) each is as highly correlated as possible with the other quotations in the field not included in the index."[2] Both requisites cannot always be fulfilled, but if the number of samples be small it is apparent that there should be a high degree of correlation of the data included with that excluded (of the same type) from the index.

Retail prices, even in the same district, are not identical for goods of the same quality. The variations would, naturally tend to be much greater for a larger geographic area. As already indicated, the Retail Price Index of the Bureau of Labor Statistics includes prices from many cities (fifty-one), which form the material for the divisional index, from which is derived the Index of Retail prices of the United States. Some idea of the method used in the securing of quotations for the food group of the index is indicated in the following quotation:

Food

Retail prices of food are secured from retail dealers through monthly reports of actual selling prices on the fifteenth of each month. The stores are selected by agents of the bureau from those patronized by wage earners. Prices are secured from every type of store—the neighborhood store, the downtown store, the department store, and the chain store. Some of these stores are credit and delivery, some cash and carry, and some cash and delivery. Effort is made to secure quotations on similar grades of commodities in the different cities. There are, however, some local customs which must be considered when any comparison

[1] Mitchell, Wesley, C.,U. S. Bureau Labor of Statistics, *Bull.* 284, p. 39.
[2] Kelley, Truman L., "Statistical Method," p. 332.

is made of the prices in the different cities. For example, the method of cutting sirloin steak in Boston, Mass., Manchester, N. H., Philadelphia, Pa., Providence, R. I., and Portland, Me., differs from that in other cities. The cut known as "sirloin" in these five cities would be known in other cities as "porterhouse." There is in these cities, owing to the methods of dividing the round from the loin, no cut that corresponds to that of sirloin in the other cities. There is also a greater amount of trimming demanded by the retail trade in these cities than in others. This is particularly true of Providence, R. I. Diagrams are given on pages 61 to 64 of this bulletin which show the methods of cutting beef in New York, Providence, Chicago, and New Orleans. In any comparison of prices in one city with those in another, due consideration should be given to the following facts:

1. The trade demands and is furnished more expensive grades of articles in some cities than in others.

2. The cities for which prices are shown in this report are widely separated; some are in localities near the source of supply, while others are a greater distance from it, making it necessary to include in the prices a greater amount for transportation.

3. Methods and costs of doing business vary greatly in different localities, due to the demands of customers and to rentals, wages, and other fixed charges or expenses.

The number of firms is apportioned according to the industrial importance of each city. At the present time for three cities, New York, Chicago, and Philadelphia, there are 30 quotations secured for each article of food, for the other larger cities, from 20 to 25 quotations, and for the smaller cities, from 10 to 15 quotations. The total number of firms quoting prices on one or more articles of food is approximately 2,000. Quite naturally firms are not constant, but when one firm drops out permanently another firm, similar in kind, is selected to replace it. Moreover, as the industrial section of a city changes, firms are added or dropped to preserve the character of price quotations.[1]

The comprehensiveness of the Wholesale Index as published by the Bureau of Labor Statistics is likewise indicated in the following tables and quotation, which indicate it to be based on the prices of 450 commodities obtained from twenty-five sources.

In the selection of commodities it has been the aim to choose only important and representative articles in each group. To this end, in addition to utilizing all available information from official sources, a careful inquiry was instituted in the principal market centers to determine which articles within the general class or group enter to the

[1] U. S. Bureau of Labor Statistics, *Bull.* 334, p. 57.

largest extent into exchange from year to year. The sources from which price quotations for the past two years have been drawn are as follows:

TABLE LXII*a*.—SOURCES OF PRICE QUOTATIONS, 1920 AND 1921

Source	Farm products	Foods	Cloths and clothing	Fuel and lighting	Metals and metal products	Building materials	Chemicals and drugs	House-furnishing goods	Miscellaneous	Total
Standard trade journals	46	80	..	10	33	26	43	..	18	256
Manufacturers or sales agents	4	19	64	9	18	13	..	21	7	155
Boards of trade, associations, etc	1	6	6	..	1	8	..	10	..	32
Federal or state bureaus	5	..	..	1	1	..	..	..	..	7
Total	56	105	70	20	53	47	43	31	25	450

So far as possible the quotations for the various commodities have been secured in their primary markets. For example, the prices quoted for live stock and most animal products as well as grains are for Chicago; wheat and flour prices are mainly for Kansas City and Minneapolis; pig iron and steel for Pittsburgh, etc. The following table shows the various markets in which wholesale price quotations were obtained:

TABLE LXII*b*.—NUMBER OF COMMODITIES OR SERIES OF QUOTATIONS, BY MARKETS, 1920 AND 1921

Market	Farm products	Foods	Cloths and clothings	Fuel and lighting	Metals and metal products	Building materials	Chemicals and drugs	House-furnishing goods	Miscellaneous	Total
Birmingham, Ala.	..	...	..	..	1	..	..	..	..	1
Boston, Mass.	1	4	..	..	..	1	..	..	6	12
Buffalo, N. Y.	..	..	..	..	..	1	..	..	..	1
Chicago, Ill.	25	22	..	3	2	5	3	4	1	65
Cincinnati, Ohio	2	4	..	2	1	3	..	..	1	13
Cleveland, Ohio	..	..	..	..	2	..	..	..	..	2
Decatur, Ill.	..	2	..	..	..	..	..	..	..	2
Gloucester, Mass.	..	1	..	..	..	..	..	..	..	1
Kansas City, Mo.	2	2	..	..	..	..	..	..	..	4
La Salle, Ill.	..	..	..	..	1	..	..	..	..	1
Los Angeles, Cal.	..	1	..	..	..	..	..	..	..	1
Louisville, Ky.	1	..	..	..	..	..	..	..	..	1
Minneapolis, Minn.	2	3	..	..	..	..	..	..	2	7
New Orleans, La.	4	3	..	..	..	..	..	..	..	7
New York, N. Y.	7	47	8	8	20	16	37	..	13	156
Norfolk, Va.	1	..	..	1	..	1	..	..	..	3
Philadelphia, Pa.	2	4	..	..	3	..	..	..	1	10
Pittsburgh, Pa.	..	..	..	2	17	..	..	..	..	19
Portland, Ore.	2	1	..	..	..	..	..	..	..	3
St. Louis, Mo.	..	3	..	1	..	1	..	..	..	5
San Francisco, Cal.	2	4	..	..	..	1	..	..	..	7
Toledo, Ohio	..	1	..	..	..	..	..	..	..	1
Wilmington, N. C.	..	..	..	..	..	1	..	..	..	1
Mills, wells, factories, etc.	1	..	..	3	6	17	2	20	1	50
General market	4	3	62	..	..	..	1	7	..	77
Total	56	105	70	20	53	47	43	31	25	450

U. S. Bureau of Labor Statistics, Bull. 320.

From these quotations it becomes evident that great care should be exercised in the selection of samples. If expense and time permitted it would be well to gather all the data, but such a task would be herculean for some studies, and far too expensive for most. Therefore, sampling must be resorted to.

All authorities agree that the samples included in an index number must be chosen with care. It is true the stress laid upon this point varies, some authorities asserting that accuracy of data takes precedence over weighting or the formula used in analysis, whereas others gives such topics equal weight.

b. Number of Samples.—Authorities disagree as to the number of samples which should be included within any one study,

although all would subscribe to Professor Fisher's statement that "certain it is that when the number of samples used is small an unwise choice can spoil the result."[1] The apparent disagreement arises in connection with the added accuracy which additional numbers of samples afford.

Wesley C. Mitchell in *Bulletin* 284 of the United States Bureau of Labor Statistics analyzes, for the period 1890 to 1913, six index numbers which comprise 262, 145, 50, 40, 25, and 25, respectively, different commodities. In the analysis he distinctly shows the 145-commodity index to be far more sensitive to the change in general business conditions than the larger index of 240 to 262 commodities. The reason, however, cannot be attributed to the number, but largely to the sample included since the two indices have different compositions. The indices of the smaller groups of commodities show greater agreement among themselves when compared to the larger indices, the reason, once more, being attributed to different samples.
He summarizes the analysis as follows:

> Large index numbers are more trustworthy for *general purposes* than small ones, not only in so far as they include more groups of related prices, but also in so far as they contain more numerous samples from each group.[2]

Professor Fisher has analyzed different distributions of some two hundred commodities chosen from *Dun's Review*. He endeavored to find the relationship between the number of samples and the accuracy of the index. His conclusions, based upon the analysis, lead him to believe that

> It is clear that the *mere number* of commodities is of only moderate importance. A small number may be nearly as good as a large number, provided they be equally well selected or assorted.[3]

This statement seems to be justified by the reliability of the Ten Commodity Index published by the Harvard Economic Service.

The statements of Wesley C. Mitchell and Professor Fisher appear to be contradictory. As a matter of fact, there is more agreement than disagreement. One is speaking largely of general-purpose indices and the other of special-purpose indices.

[1] Fisher, Irving, "The Making of Index Numbers," p. 331.
[2] Mitchell, Wesley, U. S. Bureau of Labor Statistics, *Bull.* 284, p. 58.
[3] Fisher, Irving, "Making of Index Numbers," p. 338.

Naturally, the latter has a much narrower field, so that the number could well be less.

The conclusions which may be derived from the above remarks are that in any study as many samples as time and money will permit should be chosen if they have special significance for the purpose of the index number. Likewise, the samples should be representative of all their groups.

Paul Douglas, in his article on "The Wage Studies of the National Industrial Conference Board,"[1] furnishes an excellent illustration of the discrepancies which arise in the analysis of the same subject when then are differences in type and number of samples selected. Table LXIII illustrates the differences in two published index numbers which ostensibly cover the same subject, the measurement of real wages.

TABLE LXIII.—COMPARISON OF WAGE INCREASES IN 1918 OVER 1914 AS SHOWN BY THE INVESTIGATIONS OF THE NATIONAL INDUSTRIAL CONFERENCE BOARD AND THE UNITED STATES BUREAU OF LABOR STATISTICS

Industry	Hourly earnings			Full-time weekly earnings		
	Increase as per National Industrial Conference Board	Increase as per United States Bureau of Labor Statistics	Relation of National Industrial Conference Board to Bureau of Labor Statistics	Increase as per National Industrial Conference Board	Increase as per United States Bureau of Labor Statistics	Relation of National Industrial Conference Board to Bureau of Labor Statistics
Boots and shoes....	60	39.6	+20.4	45	33.3	+11.7
Woolen and worsted manufacturing...	104	86.7	+17.3	103	85.1	+17.9
Cotton manufacturing.....	105	75.5	+29.5	106	72.6	+33.4

By Permission of Paul Douglas.

The divergence in these two indices is quite marked, and at once an explanation is necessary.

In analyzing the data comprising the two indices the article by Douglas endeavored to show that:

1. The periods covered by the two indices differ slightly.
2. The samples used differ in kind, the Bureau of Labor Index using both union and non-union wages, whereas the other index includes largely non-union data.

[1] *Amer. Stat. Pub.*, September, 1921.

3. The number of samples differed, the Bureau of Labor Index using roughly four, two, and six times the number for the respective industries of those used in the other index.

4. The samples used by the Bureau were more representative of the wage phenomena, being selected, generally, from larger geographic areas.

5. The method of obtaining the data differed.

This illustration shows the importance of studying the contents of every index number used. Both series, however, show a trend upward in real wages which is quite contrary to a study made by Paul Douglas and Miss Lamberson.[1] The following table, which has for its base the period 1890 to 1899, indicates a downward swing in real wages from 1914 to 1918:

TABLE LXIV.—INDEX OF REAL WAGES, 1890–1918

Year	Hours per week	Wages per hour	Full-time weekly wages	Retail food prices	Purchasing power, measured by retail prices of food, of	
					Wages per hour	Full-time weekly earnings
1914	90.1	153.1	137.9	158.5	96.5	87.0
1915	89.2	152.5	135.5	156.5	97.5	86.6
1916	89.7	164.5	144.8	177.6	92.6	80.8
1917	88.5	167.0	146.9	233.4	71.6	64.0
1918	88.5	211.3	187.7	266.6	79.3	70.4

By Permission of Paul Douglas.[1]

The results of the above table differ fundamentally from those in Table LXIII in that Table LXIV indicates a negative trend in real wages. The divergence can be explained largely by the type and number of samples included in the study.

In this study only union wages were recorded and the data were obtained from only nineteen industries, war-time industries, such as munition plants, being excluded. The cost-of-living figures were derived from the fluctuations in but fifteen commodities. The final wage index was really an index derived by multiplying the relative number of hours worked per week by the relative wage per hour. The resulting index, therefore, does not indicate actual wages. The final index was also derived by

[1] "The Movement of Real Wages, 1890–1918," *Amer. Econ. Rev.*, September, 1921.

the simple arithmetic mean. These are the differences which explain, in a large measure, the differences in the results of Tables LXIII and LXIV. The greatest factor, however, in explaining the difference is that the wage rate in Table LXIV was for union men only. It was well known that union rates lagged somewhat behind the non-union during the war.

These illustrations should impress upon the student the importance of the accuracy of samples and the need for sufficiency of samples in order that the study may be comprehensive and real. Finally, in using an index number made by others, the student should analyze critically its contents and purpose.

120. Weighting.—*a. Introduction.*—Weighting is the conscious effort on the part of the maker of index numbers to give weight to the data contained therein according to their relative importance.

A simple arithmetic mean gives weight to extreme items, as shown on page 233, where the price of one article doubled while that of another decreased one-half. Nevertheless, the index indicated an increase of 25 per cent. Such weighting is, of course, due to chance. The index should actually have registered 100 if the two commodities were of equal importance in the family budget. Therefore, all series are weighted, especially those having more than single frequencies, but such weighting is "haphazard" and is not the type of weighting to be considered in this section.

The importance of weighting has been minimized by some writers who emphasize accuracy of data and the correct formula in analysis in contrast to "exactness in weighting." No one would endeavor, however, to bolster up inaccurate data by means of weights; but accurate data may give a more accurate composite picture, especially if the samples are few in number, if the data have been given weights according to their importance, provided, of course that the index has been derived by the formula fitting the needs of the index. In other words, accuracy of data is of prime importance, but at the same time the data should be weighted according to their *estimated* importance.

The weights used in the construction of an index number are analogous to the frequencies attached to the class intervals of a frequency table. They indicate the number of times the given group of data should be included in the derivation of the grand total for all groups. If the index number be derived through

the arithmetic mean the group data are multiplied by their respective weights. The sum of the products is then divided by the sum of the weights.

In order to illustrate how a weighted average index differs from "haphazard" weighting, and at the same time indicate the method of computation of the weighted arithmetic mean, let the previous illustration of change in prices afford the basis of computation. Let it be supposed that article X, which increased 100 per cent in price, was one of the important articles in the family budget. This importance is indicated by the proportion expended on this article to the total cost of the food budget. Article Y, which decreased 50 per cent in price, was an article of relative minor importance in the family budget. It is self-evident with these changes in price that the increase in the amount expended by the family to maintain the same standard of comfort would certainly be greater than 25 per cent as indicated by the simple arithmetic mean of the changes.

Inversely, if commodity X was relatively unimportant, while Y was a necessity, the cost of living would have been lower when measured in terms of actual expenditure.

Granting the relative importance of articles X and Y to be 30 and 5 respectively when measured on the basis of proportional expenditure of budget, X therefore has a greater influence on the index than Y by virtue of the larger weight. Let it be assumed for the base year 1913 that articles X and Y both sold for \$1 each, but in 1924 the prices were \$2 and 50 cents respectively. The following table shows the method of computing the index.

TABLE LXV.—SHOWING DERIVATION OF WEIGHTED INDEX

Article	Weight	Actual prices		Weighted prices	
		1913	1924	1913	1924
Article X.....	30	1.00	2.00	30.00	60.00
Article Y.....	5	1.00	0.50	5.00	2.50
			Weighted aggregate value	35.00	62.50

$$\text{Weighted arithmetic mean index} = \frac{\text{Weighted aggregate}}{\text{Total weights}}$$

$$\text{Index for 1913} = \frac{35}{35} = 100.$$

$$\text{Index for 1924} = \frac{62.50}{35} = 178.$$

The simple arithmetic mean indicated for the two articles a price increase of only 25 per cent, whereas the weighted arithmetic mean, since the index is 178, indicates an increase of 78 per cent, which shows that to maintain the same standard of living for food, the family income should be increased 78 per cent. Reverse the weights, that is, assign an importance of 30 to Y and 5 to X and the weighted index for 1924 becomes $25\frac{5}{35} =$ 71 per cent, which indicates a considerable decrease in the cost of living. This illustration should serve as an example of how the weights affect the results, and furthermore emphasize the need for both accuracy of data and reliability of weights. Weighting is particularly necessary for index numbers when computed from relatively few samples, whereas in larger indices weights may be given by including some items in their different forms (raw materials in their finished goods). In any case weighting should be given due consideration in order to avoid haphazard results.

Weighting, however, does not always involve the division of the sum of the products of weights and attributes by the total weights unless the index be derived by means of averages. It is possible to leave the data in absolute numbers in the form of weighted aggregative as Professor Fisher has named it. For the above illustration the base year could be represented as an aggregate figure of $35 and the year 1924 by $62.50. If the weights have remained constant during the period, it is possible to reduce these absolute values to relative figures on any desired base. This may be accomplished by dividing the aggregative value of any other years by the value for the desired base year. If 1924 were selected as base the year 1913 would have a relative of $\frac{35}{62.50}$, or 56 per cent, whereas, as already found, the 1913 base yields an index of 178. This is a decided advantage in the use of the aggregative, which will be discussed more fully later.

b. The System of Weights to Use.—If weighting is important it may be asked: What system of weights should be used? There is no universal rule for the choice of the weights or the period from which they should be derived. The system adopted, once again, generally, depends upon the purpose of the index. Therefore, the weights used in various indices are not interchangeable unless the purpose is identical.

The weights used in the Retail Price Index of the United States Bureau of Labor Statistics are essentially cost-of-living weights. The weights used in the subindex, food, are given in Table LX, page 235, which indicates the system to to based upon the yearly quantity of the various articles consumed by a family of five. Naturally, these weights are useless and meaningless when applied to other price phenomena, such as wholesale prices, farm-crop prices, etc. For the weights used in the construction of the Wholesale Index as published by the same department the student is referred to the appendix of *Bulletin* 335.

Weights are devised according to the purpose of the study and some idea of the systems used may be gathered from the following references to several indices.

Wholesale Prices.—The average price of each article in the year 1913 (selected as the base period in order to provide a convenient pre-war standard for measuring price changes) has been multiplied by the estimated quantity of that article marketed in the census year 1919.[1]

Physical Volume of Production.—Agricultural Index.[2]—The several commodities comprised in this index will be weighted according to their respective economic importance as evidenced in the average annual value of the individual crops during the base period 1909 to 1913.

Combined Index.—While the method of weighting agriculture, mining and manufacture according to aggregate values in 1909 is quite consistent with the method of weighting employed elsewhere in the study, and is presumably preferable at this point in the analysis to any other method of weighting, it is interesting to note the effect upon the combined index when the three components are weighted, not according to aggregate values, but according to the number of persons engaged, as reported in the census count of 1910. The numbers engaged in agriculture, mining, and manufacture respectively in that year were 12,384,000, 977,000, 8,101,000, and the corresponding weights to be employed in deriving the combined index are consequently 57, 5 and 38 . . . It will be observed that the indices obtained by the two systems of weighting are very much the same, despite the substantially different weights assigned to agriculture and manufacture under the two methods. *The results are merely another illustration of the stability of index numbers under varying systems of logical weighting.* (Italics not in quotation.) Both methods of weighting the combined index are logical. But the method of weighting according to aggregate values seems more consist-

[1] U. S. Bureau of Labor Statistics, *Bull.* 335.

[2] Day, E. E., "An Index of Physical Volume of Production," *Rev. Econ. Stat.*, p. 247, September, 1920.

ent with the significant relationships between agriculture, mining, and manufacture and has the merit of following the precedent set in the combinations of other series in the present study.[1]

New Index of Volume of Trade.[2]—As far as the material was available the weights assigned were according to the usual procedure; first, of comparing either the value added in manufacture, or the value of the product, or the value of the commodities exchanged; these data were checked by the figures for employment where possible.

Wholesale Prices, 1866 to 1899.[3]—The weights consist of the domestic production marketed in the United States in 1879 or 1880 plus amounts imported.

Retail Price Index.[4]—The average money price of each article for each month has been multiplied by the amount of the article consumed annually by the average family.

From these quotations it is seen that weights may be a combination of quantities or values of physical production, or quantity or value of goods exchanged. The quotation relating to the "Combined Index of Physical Production," emphasizes the fact that there are many ways by which the series of an index may be weighted, and the systems adopted matter little, provided they have a *logical basis.* Therefore, a value series should never have weights in the terms of value, since such weighting would be most illogical.

All of the above systems of weighting adopt *constant* weights; that is, the weights are fixed for a period of time and do not vary for any year of the study. The periods used as the bases for weights, however, differ, some using more remote dates, while one uses a comparatively recent date for the derivation of weights. Constant weights, however, if the period used in determining the weights is remote may distort the index. It is a true saying "what were luxuries in the family budget some fifty years ago are now necessities." Some commodities have, therefore, ceased to occupy a place in the budget, while new ones have taken their place. Clearly for such index the weights should be constantly revised. Likewise this is true for any study which involves rapidly changing conditions. Professor Fisher,

[1] *Ibid.*, p. 19, January, 1921.

[2] Snyder, Carl, "A New Index of Volume of Trade," *Amer. Stat. Pub.*, p. 953, December, 1923.

[3] Snider, Joseph, "Wholesale Prices in the United States." *Rev. of Econ. Stat.*, p. 107, April, 1924.

[4] U. S. Bureau of Labor Statistics," *Bull.* 334.

by means of his "ideal formula," permits the changing of weights annually. These are the two extremes, and perhaps a middle ground is better (unless it be a purchasing power index)—to have a fixed base for the derivation of weights—but the base should be revised when any decided change takes place in the required composition of the index. The United States Bureau of Labor Statistics at the inception of its Retail Index intimated that it would be revised at least every ten years, and in 1918 the system of weighting was revised.

c. Summary.—Summarizing this section it may be said:

> If the construction of an index number be undertaken one of the most important questions to be settled is that of logical weights to use; logical that is, with respect to the purpose of the proposed index.[1]

Weighting, therefore, must receive the careful attention of the index maker, but it should not detract from the search for accuracy of data. A well-defined purpose helps considerably in the choice of weights. Constant weights are usually used, but they should be revised with any marked changed condition. Weights may take the following forms:

1. Production rates—quantity or values.
2. Exchange rates—quantity or values.
3. Quantities entering into consumption.
4. Aggregate expenditure—proportional expense.

And these weights may be taken from

1. The base year.
2. Any given year.

121. Summary of Chapter.—This chapter may be briefly summarized as follows:

1. An index number is a number which characterizes a composite group, and may be expressed as a relative to a base which has a value of 100, or as an aggregative.

2. Since index numbers are used to measure changes in specific kinds of data, the types of data included within the index should conform to the purpose of the index.

3. Since the index is a figure which represents the values of all the types of data contained within the index, and, furthermore, is supposed to represent all like data not included within the index, it should be derived from samples which are strictly repre-

[1] PERSONS, WARREN M., "Fisher's Formula for Index Numbers," *The Rev. Econ. Stat.*, p. 111, May, 1921.

sentative of the groups. This fact may necessitate the inclusion of many samples.

4. All data are not of equal importance and, since the index should be representative of all data, the type of data should affect the index according to its importance. This may be done by a conscious system of weighting, the assigning of weights relative to the importance of the various data.

5. Weights may be derived from a fixed period, the weights remaining constant throughout the study, or may vary from year to year. A fixed system of weights should be made flexible by revision as conditions change.

CHAPTER XXI

INDEX NUMBERS (*Continued*)

I. CHAIN RELATIVES AND FIXED-BASE INDICES

122. Introduction.—With decision as to types of data, kind and number of samples, and weights to be employed in the construction of an index number, the raw materials may now be gathered for the index.

After collection, the next thing required is decision as to the form of index numbers, whether they are to be expressed as relatives or in absolute numbers called aggregates. If relative numbers are used, a point of reference is required from which to measure changes. This standard may be a definite period, called a fixed base, or it may change from year to year, forming a chain of relatives. The differences between these two methods are illustrated in Table LXVI.

TABLE LXVI.—WHOLESALE PRICES IN THE UNITED STATES
Source: Bureau of Labor Statistics

Bases		Years						
		1913	1914	1915	1916	1917	1918	1919
1913 as fixed base		100	100	101	124	176	196	212
Chain relatives	1914 as base for 1915			101				
	1915 as base for 1916				123			
	1916 as base for 1917					143		
	1917 as base for 1918						111.3	108
	1918 as base for 1919							
Continuous chain		100	100	101	124	177.3	197.3	213

Fixed Base.—With a definite year assigned as the base year it becomes apparent that the relatives for the other years are

derived by dividing the sum of the relatives, or aggregate values for each year by the sum for the base year, which year has been given a value of 100. Also with a fixed base the point of reference is but one date, 1913. The table indicates that wholesale prices in 1919 had increased 112 per cent since 1913.

Chain Relative.—The chain index, on the other hand, indicates yearly percentage change, each preceding year serving as the basis for computation. Table LXVI indicates wholesale prices in 1917 to be 43 per cent higher than in 1916. Such an index measures short-time change and does not indicate comparison of the present with the more remote.

Continuous-chain Relatives.—Chain relatives, however, may be referred to a fixed base by multiplying successively the chain relatives, that is, the 1913 relative, by 1914, 1915, 1916, 1917, 1918, and 1919 chain relatives. This operation results in a continuous chain which is linked to the 1913 base. Such a series of indices are shown in the last row of Table LXVI, their derivation being indicated by the lines.

123. Chain Relatives.—In comparing the indices derived from a fixed base and those from a continuous chain (when the arithmetic mean is used in computing the index), it should be observed that the chain yields a value of 213 for 1919 on the 1913 base, whereas, by actual computation from the fixed base, the index for 1919 was found to be 212. Since both are referred to a fixed base 1913, it would be presumed they should be identical. This difference, however, proves to be a defect in the continuous-chain system since chain relatives "usually rise more than the base relatives when prices are rising and less when prices are falling."[1] In this illustration prices were rising. Hence the chain index tends to be higher than it should be. Over a long period of time, it can be readily seen how these discrepancies become accumulative and thus distort the index.

Continuous-chain relatives should be used with considerable care, since they are difficult to interpret and do not really express the actual percentage change from the fixed base.

For comparing year-to-year changes, the chain relative has a decided advantage over the fixed base, since the yearly changes in base permit easy revision of weights and commodities, whereas the fixed base does not indicate, at once, yearly changes or permit ease in revision of the index.

[1] Mitchell, Wesley C., U. S. Bureau of Labor Statistics, *Bull.* 284, p. 87.

It may appear from the above that when indices are expressed in relative numbers the fixed-base system is advocated in preference to the chain-relative and, at the same time, the chain-relative method is preferable to the fixed-base. These statements are not contradictory, if it is borne in mind that index numbers are constructed for diverse purposes. If long-time change is desired the fixed-base system should be used, whereas for annual changes the chain index is preferred. Therefore, for general-purpose indices it would seem logical for them to be expressed in both systems.

Since the fixed-base system is used to a considerable extent in the construction of indices, it will be discussed more fully than the chain relative.

124. The Fixed Base.—*a. The Length of the Base Period.*—As already indicated the base period is represented by 100. In a study, however, there may be no year which actually indicates 100, since the period used as a base may be a series of years. The average of the sum of the data for this period is then taken as representative of 100. No matter the period used, the sum of the measurements, either relative or aggregate, for every year is divided by the sum representative of the base period, and since the base period has been given a value of 100, the quotient should be multiplied by 100, resulting in the index for the given year.

Naturally, changes are usually measured from the normal, and since the base period is the point of reference for the index this point should usually represent the normal. This would usually preclude any short period, as one month, serving as a base. Any time element, however, may serve as a base, the time chosen depending upon the purpose of the index. Since there are seasonal and cyclical fluctuations, any index which measures from the normal should rarely use a period of one year or less as base. The Index of Physical Volume of Production, as derived by E. E. Day,[1] uses for the base period the years 1909 to 1913.

The length of the base period used, therefore, depends upon the purpose of the index. If it is to measure changes from normal, it should be such as to allow for seasonal and cyclical changes.

b. The Time of the Period Chosen.—While it is possible to have represented as the base period one of sufficient length to

[1] *Rev. Econ. Stat.*, September, 1920.

indicate the normal at the time of construction, nevertheless with the lapse of time the so-called normal may become abnormal; in other words, conditions have changed.

Naturally, the more recent the base period the more likely are the present conditions to be homogeneous with those of the base period, and, inversely, the more remote the less homogeneous. A cost-of-living index based on the year 1850 would be meaningless as a point of reference for measuring the recent changes in the cost of living. Other things are rarely equal and, with time, tastes change and these change the constituent parts of the budget. Therefore, if the object of the index is to measure long-time changes with special reference to the present the base chosen should be as near in time to the present as the length of the base period will permit, provided, of course, that the present is normal.

A systematic system of weighting adopted for an index may be vitiated if the base chosen is very remote, and if the indices are computed by the arithmetic mean. The mean gives weight to extreme items and in the course of time prices and demand disperse irregularly from the base-period relatives. Some group-relatives, therefore, which make up the index may give undue influence to the final index because of dispersion in price from the normal, thus vitiating the weights given to the various groups of data because the mean emphasized the abnormal.

Even presuming conditions of demand were identical in 1920 and in 1850, of what significance would the cost-of-living index for 1920 on the 1850 base be to the workman who is trying to maintain his standard of living by wage demands? Clearly, the more recent figures are necessary, and in this illustration the chain index would be preferable to the fixed-base index.

The fixed-base system has its defects. Therefore, the period used as base should be homogeneous, as far as possible, with the conditions which the index aims to measure.

c. The Difficulties of Changing Base or Commodities in a Fixed-base Index.—With a fixed base all indices are expressed as relatives to the base, and therefore are measured from one definite point of reference. Changes from that point are readily recognized by the divergence in the respective indices. It may be necessary, however, to derive percentage changes of one point of time to another; in other words, it may be necessary to shift the point of reference either temporarily or permanently.

The problem may necessitate the derivation of the change in wholesale prices from February, 1924, to March, 1924. On the 1913 base the respective relatives are 152 and 150, the March relative indicating a decline of two points. This is not a 2 per cent decrease, since it is a difference from the February relative which has a point of reference in 1913. The decrease, however, may be expressed as a percentage change from February by considering the February relative to be the new base. March, therefore, has a relative of $^{150}/_{152}$, or 98.6 when compared to the February relative *on the* 1913 *base*. The base has been shifted to February, 1924, but it is a base which is still linked indirectly to 1913. For short periods of time, however, this method may be adopted in changing the base.

If, on the other hand, the base was to be shifted some twenty years, from 1900 to 1920, and the 1920 relative given a value of 100, while all antecedent and subsequent yearly relatives were divided by the 1920 relative on the 1900 base, the results would not be accurate. In twenty years the prices or quantities making up the 1920 relative have diverged in different degrees and direction from those comprising the 1900 base relative. The 1920 relative was derived by (assuming the arithmetic mean is used) obtaining the mean of the relatives for the groups which comprise the index. The mean is an average which gives weight to extremes. Therefore with the 1920 relative as base some groups may be given, unconsciously, more weight in deriving the index than others. Wesley C. Mitchell has, in *Bulletin* 284, page 85, analyzed the discrepancies that arise by such a process of base shifting. In a fixed-base index, if accuracy is desired, the base can only be changed by a recomputation of all price relatives to the new base. By using the geometric mean such a process may be avoided, since the relatives so derived may be shifted to any base at will.

Table LXVII illustrates the above discussion. In this table commodities X and Y constitute the simple unweighted index; the actual values of these commodities are given for the old base year and the required new. The actual prices are shown in columns A and B, while their relatives are indicated in C and D respectively.

In this table the mean for the 1924 prices on the 1913 base is found to be 112.5 (row 2). If the method of changing bases as

TABLE LXVII.—TABLE SHOWING THE DEFECTS OF BASE SHIFTING BY THE RELATIVE METHOD, UNLESS COMPUTED BY THE GEOMETRIC MEAN

Year	Actual prices		Relative prices		Arithmetic index $\frac{C+D}{2}$	Geometric index $\sqrt{C \times D}$
	X	Y	X	Y		
	A	B	C	D	E	F
1. Base 1913	1.00	1.00	100	100	100	100
2. 1924	2.00	0.25	200	25	112.5	70.7
3. Base 1924	2.00	0.25	100	100	100	100
4. 1913	1.00	1.00	50	400	225	141.4

discussed above is used in this illustration that is, assuming the 1924 relative of 112.5 to equal 100, the 1913 relative (100) on the 1924 base would equal $\frac{100}{112.5}$, or 88. Such a figure is not the actual relative as shown by rows 3 and 4, in which the relative for 1913 on the 1924 base was computed from the actual data and resulted in a relative of 225 for 1913 on the 1924 base. There is quite a divergence from the relative method for 1913, (88) and the actual recomputation, (225). This extraordinary divergence is explained in the use of the arithmetic mean, which gives weight to extremes. This illustration indicates the inaccuracies that may arise by arbitrarily assuming the relative for any desired period, as the new base, to be equal to 100. It also emphasizes the difficulties of shifting bases when the index is computed by the arithmetic mean of price relatives.

Column F of the above table indicates how, by the use of the geometric mean, relatives may be shifted readily to any base by the division of other relatives by that chosen for the base. The geometric mean for the 1924 values on the 1913 base was found to be 70.7 (row 2). If the 1924 relative on the 1913 base be chosen as the new base; the 1913 relative on the 1924 base would be $\frac{100}{70.0}$, or a value of 141.4. This value is identical with that derived by actual recomputation of the index as shown in rows 3 and 4. Therefore, an index derived by the use of the geometric

mean may be shifted readily to any desired base without recomputation of the relatives comprising the index. This mean, however, is not generally used, owing to its complexity of computation. Therefore, most index numbers, when computed from a fixed base, require recomputation of all relatives of the groups if the base is shifted.

The attention of the reader might be called to an apparent incongruity of the arithmetic mean. Column E, Table LXVII indicates a 12.5 per cent increase of the 1924 prices over those for 1913, also a 125 per cent increase of the 1913 prices over the 1924 base. There is no real inconsistency in these statements, since two different bases are used to derive the figures. The relative increase from one base may be a decrease when measured from the other base. The geometric mean, on the other hand, since a 400 per cent increase in price of one commodity is offset by a 75 per cent decrease in another, irrespective of the quotations, does not show such apparent inconsistencies. The relatives for the 1913 prices are higher than those of 1924 irrespective of the base.

Another problem presents itself in a fixed-base index when the index has to be revised to include new and exclude some old commodities. The time has been determined from which to measure changes, but with the passing of time some commodities become obsolete while new demands create new products. The new products may be of such importance as to warrant their inclusion in the index. Naturally, if ten (10) years have elapsed from the base period of the index it would be folly to embody in the index the new commodities at their current prices. The point of reference is ten years past and therefore some basis of computing a relative of the new commodity which will show its change from the old base period is necessary. It is a simple but laborious process if quotations exist for the commodities from the base period, for relatives can easily be constructed for the previous ten years. Sometimes, however, such quotations are not obtainable; a new method of deriving the relative for the current year on the old base for a new commodity has to be devised. This can perhaps be best illustrated by the method used in the price indices as published by the United States Bureau of Labor Statistics.

Price quotations extending back to the base period 1890 to 1899 were unobtainable for the newly introduced articles. The adoption of some

method of calculating hypothetical relative prices for the new commodities was necessary if the new commodities were to be introduced into the index number. The method adopted by the Bureau may best be made clear by describing how the index number of a particular group was computed. Let us consider the farm-products group. In this group horses, mules, live poultry, and Burley tobacco were included for the first time in 1908. Prices of these new articles were obtained for both 1907 and 1908. A relative price for each of the 20 old and new articles included in the group was calculated for 1908 on the 1907 base. These relative prices were added together and divided by 20, the number of commodities in the group, to get the simple arithmetic average of the relative prices of farm products in 1908 on the 1907 base. This group index number was then multiplied by the 1907 index number computed on the money prices of the 16 old articles to obtain the 1908 index number of farm products on the 1890 to 1899 base . . . The method can be illustrated just as well by the computation for 1909 given below.

Relative Prices of Commodities in 1909 Computed on Actual Prices in 1908, as Base

Commodity	1909 relative price on 1908 base	Commodity	1909 relative price on 1908 base
Cotton	115.7	Steers, choice	107.6
Flaxseed	130.2	Steers, good	107.7
Barley	91.9	Hogs, heavy	130.6
Corn	97.6	Hogs, light	130.6
Oats	94.4	Wethers, native	109.7
Rye	100.0	Wethers, western	109.5
Wheat	121.2	Poultry	120.3
Hay	109.1	Tobacco	116.8
Hides	123.3		
Hops	169.0	Total	2,299.7
Horses	103.6	Average	114.99
Mules	110.9		

The group index number for 1909 on the 1908 base (114.99), multiplied by the group index number for 1908 on the 1890 to 1899 base (133.1), gives 153.1, which was taken as the index number for farm products for 1909 on the 1890 to 1899 base.

[1] U. S. Bureau of Labor Statistics, *Bull.* 181, p. 244.

d. Summary.—From the discussion of the fixed base it becomes apparent there are difficulties (1) in interpreting the yearly indices; (2) in maintaining homogeneity of conditions; and (3) in the including new commodities. The difficulty of obtaining yearly comparisons may be overcome by resorting to the chain relative, by expressing each yearly index as a percentage of the preceding one. In this way yearly comparisons are made as shown in Table LXVI. The difficulty of shifting bases may be avoided if the geometric mean is used in deriving the relatives. There is no easy path to the inclusion of new commodities.

Because of the complexity of the geometric mean, the unreliability of the chain index when computed from the arithmetic mean, and the difficulty of adding new products, the aggregate of actual values may serve as a basis from which indices may be constructed, at will, on any base and at the same time, permit easy inclusion of new commodities.

II. THE AGGREGATIVE

125. Introduction.—Since index numbers expressed as relatives to a fixed base have certain defects, it is possible to express indices in the form of actual sums called "aggregatives," which in price series is the total cost of a bill of goods at different periods of time. These sums may be made relative numbers by dividing them by the aggregative for the desired base period. The indices thus become relatives. In some indices the relatives, as derived above, for the groups are then averaged, as in the Wholesale Price index of the Bureau of Labor.

In this wholesale index the aggregate method is used in deriving the price relatives for each of the nine groups comprising the index. These nine relatives are then averaged. More specifically, each of the nine groups has an index which is derived by multiplying for each article within the group, the price quotation, for the current period, by the quantity of the commodity exchanged in the year 1909. These products when summated yield the cost of a bill of goods for the group. The cost when divided by the cost in 1913 gives the relative price for the group on the 1913 base. This process is continued for each group. The combined index is then derived by averaging the group relatives computed from their aggregates.

126. Aggregates Shifted to Any Base.—The advantage of the aggregate method in contrast to averaging relative prices for

any one group and then averaging the group averages may be illustrated by the following table, although it deals with only one commodity:

TABLE XLVII*a*.—DIFFERING RESULTS OBTAINED BY SHIFTING BASE PERIOD OF RELATIVE PRICES COMPUTED BY OLD AND NEW METHODS[1]

Round Steak. (*An example of an article whose prices fluctuate rather narrowly*)

Firm	May		June			July		
	Price	Relative on May base	Price	Relative on May base	Relative on June base	Price	Relative on June base	Relative on May base
A	B	C	D	E	F	G	H	I
3002	$0.23	100	$0.25	108.70	100	$0.25	100.00	108.70
3013	0.25	100	0.25	100.00	100	0.25	100.00	100.00
3012	0.23	100	0.23	100.00	100	0.25	108.70	108.70
3017	0.24	100	0.22	91.67	100	0.24	109.09	100.00
3018	0.23	100	0.22	95.65	100	0.25	113.64	108.70
3023	0.22	100	0.25	113.64	100	0.25	100.00	113.64
City aggregates	1.40	600	1.42	609.66	600	1.49	631.43	639.74
City relatives—Averages of firm relatives	..	100	..	101.61	.....	..	105.24	106.62
City relatives computed from actual prices, *i.e.*, $1.42 ÷ $1.40, $1.49 ÷ $1.42, and $1.49 ÷ $1.40	.....	100	.....	101.43	100	.....	104.93	106.43

City relative for July on May base computed by averaging relatives and multiplying the averages, *i.e.*, 101.61 × 105.24 = ... 106.93

City relative for July on May base computed by multiplying relatives computed from aggregate actual prices, *i.e.*, 101.43 × 104.93 = ... 106.43

The table indicates the relative price, 106.62 (column I)—obtained by averaging the July price relative referred *directly* to the May base is 0.19 point greater than the relative, 106.43, derived by dividing the *actual aggregate* figures for July, $1.49, by those for May, $1.40. Also, the relative for July on the May base when derived by the continuous chain of relatives yields an index of 106.93, or 0.31 point greater than the direct computation of the July relative on the May base, whereas the continuous chain relative for those derived from the *aggregate* prices yields a July relative on the May base identical with that derived by direct computation from the aggregate figures $\left(\frac{1.49}{1.40}\right) = 106.43$.

[1] U. S. Bureau of Labor Statistics," *Bull.* 156, p. 367.

Therefore, relative prices computed by comparing aggregate actual prices can be constructed month by month and shifted to any desired base period, year or month, with mathematical accuracy *so long as the actual price quotations come from identical firms throughout the period considered.* This is readily shown by the following formulae: $\frac{\$1.42 = \text{June aggregate price}}{\$1.40 = \text{May aggregate price}} = 101.43$, the June relative price on the May base. $\frac{\$1.49 = \text{July aggregate price}}{\$1.42 = \text{June aggregate price}} = 104.93$, the July relative price on the June base. Multiplying the July relative price on the June base by the June relative price on the May base, we have $104.93 \times 101.43 = 106.43$, or $\frac{\$1.42}{\$1.40} \times \frac{\$1.49}{\$1.42} = \frac{\$1.49}{\$1.40} = 106.43$, the July relative price on the May base, which is exactly the same result as would be obtained by dividing the July aggregate in the first instance by the May aggregate. Even with changing firms and commodities varying in quality, the relative prices calculated by comparing actual prices may be shifted to any desired base by the method illustrated above with the closest possible approach to absolute accuracy.[1]

The aggregative, therefore, has a decided advantage over the fixed base in that it may readily be changed to any base.

127. Weighting an Aggregative.—An aggregative, however, is not so simple in construction as indicated in the table above, for index numbers measure group changes and not those in single commodities. Groups and commodities within groups have different units of measurement and relative degrees of importance. The choice of prices to include within a group is but a preliminary step, since they must be weighted. Weighting is of extreme importance in aggregatives.

If an index number comprised the mere summation of prices, the articles usually quoted by the ton would constitute a greater part of the aggregate sum than those sold by the pound, if the costly articles were excluded from the index. Such an index, therefore, would be meaningless in measuring group changes. Some system of weighting is essential. The Bradstreet Index reduces its ninety-six articles to a cost per pound (avoirdupois), thus making the quantitative weight identical for all. A system such as this would give fair results only when the index comprises articles having a fair degree of correspondence in their relative importance. If such were not the case a vicious system of

[1] *Ibid.*, p. 369.

TABLE LXVIII.—ILLUSTRATIVE TABLE OF INDEX NUMBERS COMPUTED FROM PRICES OF FIVE COMMODITIES[1]

Commodity	Unit	Price		Price per pound (avoirdupois)		Relative price 1913 on 1912 base	Relative price 1912 on 1913 base	Amount exchanged in 1909 (000 omitted) used as weight	Value in exchange (000 omitted)	
		1912	1913	1912	1913				1912	1913
A	B	C	D	E	F	G	H	I	J	K
Silver	Ounce, troy	$ 0.61468	$ 0.60448	$ 8.96418	$ 8.81533	98	102	151,969 ounces	$ 93,412	$ 91,862
Opium	Pound, apothecary	7.1167	6.0167	8.64876	7.31196	85	118	491 pounds	3,494	2,954
Coal (bituminous)	Ton, long	3.1258	3.0542	0.00140	0.00136	98	102	288,083 long tons	900,490	879,863
Petroleum (crude)	Barrel, 42 gallons	1.5750	2.4500	0.00469	0.00729	156	64	183,171 barrels	288,494	448,769
Pig iron	Ton, long	14.2396	14.0979	0.00635	0.00629	99	101	8,025 long tons	114,273	113,136
Total		$26.67178	$26.22328	$17.62538	$16.14223	536	487	.....	$1,400,163	$1,536,584
Index numbers for 1913 calculated on 1912 prices as 100		98		92		107	...	.....	110	
Reciprocals of 1913 indexes		102		109		93	...	.....	91	
Index numbers for 1912 calculated on 1913 prices as 100		102		109		...	97	.....	91	

[1] U. S. Bureau of Labor Statistics, *Bull.* 181, pp. 248 and 249.

weighting would be introduced, since too great a weight would be given to the price figure of the costly, and perhaps relatively unimportant, article, and too little to the low-priced necessities. An excellent illustration of this need for weighting is afforded in Table LXVIII.

The absurdity of an index derived from the summation of mere prices is shown in columns C and D. These sums, as it should be observed, are derived from diverse units of measurements, and yield an index of 98. Columns E and F yield an index of 92 on the 1912 base. This index gives the same quantitative weight of 1 pound to all commodities, with the result that the decline in price of silver and especially opium in 1913 had more influence on the aggregate value when quoted by the pound than in the diverse weights of columns B, as evidenced by the fall in the index from 98 to 92. In spite of this smaller index, petroleum actually increased in value in 1913. Column G illustrates the index derived by averaging price relatives for 1913 on the 1912 base. This index of 107 is influenced greatly by the relative increase in the petroleum prices, which has a relative of 156. This indicates, once more, that the averaging of relatives gives weight to extremes. Column I assigns weights to the articles according to the "amount exchanged in 1909," and by the use of these weights the index for 1913 on the 1912 base becomes 110, which is the correct index and is derived by dividing the aggregate $1,536,584 for 1913 by $1,400,163 for a 1912.

This aggregate index, it is seen, is easy of computation and may be readily be shifted to any desired base, comparability of all and any years being possible. Also, when the aggregates are expressed as relatives they may likewise be changed to any base. The important feature of an aggregate as emphasized by the above is a logical system of weights. Neglect of weighting vitiates accuracy of computation.

128. Summary on Aggregates.—An aggregate may be expressed in absolute numbers or expressed as a relative of the aggregate sum either to a fixed base or as a chain relative. These chain relatives when linked together produce relatives identical with those computed from the actual aggregates. In other words, aggregates may be shifted to any desired base, Table LXIX illustrates this summary.

The last two rows of this table, it should be observed, are identical, though derived from different sources.

TABLE LXIX.—SHOWING HOW CONTINUOUS-CHAIN RELATIVES YIELD IDENTICAL RESULTS WITH ACTUAL COMPUTATIONS, ALSO THE POSSIBILITY OF SHIFTING BASES

Index of X

		1913	1914	1915	1916	1917
Aggregate values		$2.00	$3.00	$4.50	$5.85	$11.70
Relatives	1913 as base for 1924	...	150			
	1914 as base for 1915	...	...	150		
	1915 as base for 1916	...	...	...	130	
	1916 as base for 1917	...	...	...	...	200
	Continuous chain	100	150	225	292.5	585
	From aggregate figures	100	150	225	292.5	585

Since the aggregative is a sum of actual prices or quantities it should of necessity have weights in order to avoid the vicious weighting which could arise through differences in actual quotations which are not always criteria of the relative importance of commodities.

The aggregative is easily computed and readily understood and lends itself to further algebraic treatment. Furthermore, since it is expressed as a sum or quantity, the index may readily be revised, it being relatively simple to add or subtract commodities.

It is the form of index which should be used to measure, not ratios of change, but the cost of a bill of goods.

Wesley Mitchel in speaking of aggregates states:

> In fact, they combine most of the merits and few of the defects characteristic of the various methods of averaging relative prices.[1]

III. AVERAGES TO USE

129. Introduction.—Since index numbers may take the form of relatives related to a fixed base, chain relatives, and aggregatives, it becomes apparent that the computation of some demands the use of averages in order to derive the index.

[1] U. S. Bureau of Labor Statistics, *Bull.* 284, p. 80.

130. Aggregatives May Be Devoid of Averages.—An aggregative may be expressed as an absolute number, or this may be reduced to a relative on any desired base.

An index number usually gives a composite picture of several groups. If the total, or aggregate, for each group is multiplied by its relative importance and the products for all groups summated, a total or aggregative is reached which serves as an index number for the combined groups. Such an index has been derived without the use of any average and in its form of absolute numbers it may be referred to any base and thus possesses the possibility of comparability with any series on any base (provided the series cover the same period). The aggregative, therefore, has many advantages over types which employ averages, since all averages have their limitations, there being no ideal average for all purposes.

It is true, however, that some aggregative index numbers average the relatives of the aggregates for each of the groups and thus give a combined index. Such an index is usually obtained by deriving the arithmetic mean of the group aggregate relatives.

The aggregative form of index is the most satisfactory if changes in the cost of a given bill of goods is desired. If, however, ratio changes in prices or quantities are required, the geometric mean should be employed in the derivation of an index.

131. The Geometric Mean.—The geometric mean, as already indicated in Table LXVII, when applied to data with a fixed base, is such that the base may be readily shifted to any year without the need for recomputation of all relatives.

The geometric mean measures ratio of change. One commodity which doubles in any index has precisely the same weight in the determination of the index as the halving of another; that is, if one price or quantity increases from 200 to 400 and another falls from 2 to 1, the commodities being of equal importance and the other items in the index remaining unchanged, the index remains unchanged, since $\sqrt{200 \times 2} = \sqrt{400 \times 1}$. This type of average should, therefore, be used when ratio of change is important and *never* when undue emphasis is to be given to extreme items.

The geometric mean is used in the index measuring the Physical Volume of Production.

This mean is adopted for two reasons: (1) it reduces somewhat the influence of extreme items—a considerable advantage in constructing a fixed-base index; and (2) it makes it possible to shift the base of the index freely without recalculation of the averages—a possibility which will be valuable in the later parts of the present study.[1]

From the above quotation it is evident that the geometric mean is always smaller than the arithmetic mean, since it does not give such weight to extreme items.

This mean, because of the complexity of computation, is not generally used. It should be used, however, whenever the purpose of the index warrants such an average.

132. Median and Mode.—If the purpose of an index number were to ignore extreme items, the type of average to use in the derivation of the index would be either the median or mode, since both are derived without measurements of extremes.

These averages, however, have their limitations. A series may be bimodal and if the data are scattered and few in number the median may hardly be representative.

Irving Fisher in speaking of averages to use in the construction of index numbers states:

> The mode and, in general, the median may be ignored as too sluggish or unresponsive to small influences to make them sensitive and accurate barometers of price movements.[2]

The median and the mode should be used sparingly in the determination of indices and only when purpose requires no emphasis of extremes and when the data are sufficient in quantity and concentrate around a norm.

133. The Arithmetic Mean.—The arithmetic mean is most commonly used in general-purpose indices, its ease of computation and comprehension appealing to the maker and user of index numbers.

As already emphasized, this mean gives weight to extreme items. Its ease of computation is apparent when contrasted to the geometric mean. The latter, however, has the advantage of base shifting without recomputation.

[1] Day, E. E., "An Index of the Physical Volume of Production," *Rev. Econ. Stat.*, note, p. 247, September, 1920.

[2] Fisher, Irving, "The Making of Index Numbers," p. 351.

Ease of computation should not be the sole criterion in the determination of the average to use, it must always conform to *purpose.*

134. Averages to Use.—Wesley Mitchell gives an excellent summary of the averages to use when he states:

Wise choice of the average to use in making an index number, then, involves careful consideration of the materials to be dealt with and of the purpose in view. (1) If that purpose be to measure the *average ratio of change* in prices, the geometric mean is the best; indeed, in strictness, it is the only proper average to employ—on one interpretation of that somewhat indefinite problem. For alone among our averages the geometric mean always allows equal influences to equal ratios of change price, quite irrespective of the previous levels of the prices in question, the amounts of money represented by the changes themselves, or any other factor. As has been said already, in a geometric mean the doubling of one price is precisely offset by the halving of another price —though if the two prices were originally the same the rise amounts in money to twice the fall. And further changes of 10 per cent from the two new prices will again be precisely equal in their influence upon the geometric mean, although 10 per cent of the price that has doubled represents a sum of money four times as great as 10 per cent of the price that has been halved. (2) But these same examples show that geometric means are not proper averages for measuring alterations in the amount of money that a given bill of goods costs. And as a rule, our interest does center in the money cost of goods rather than in the average ratio of changes in price. For example, when we are investigating the increased cost of living, the doubling of one item in the family budget may well be twice as important as its halving; and when we are studying the "relation of prices to the currency, a large upward variation should count for more than a small downward variation, for it requires more currency;" provided always that the changes in prices are not offset or more than offset by contrary changes in quantities bought. For such purposes the arithmetic mean is the logical average to use. (3) Frequently, however, the very fact that an article has advanced greatly in price cuts down it market, so that the increase in money cost represented by the arithmetic mean exists on paper rather than in fact. When such cases of extreme advance are numerous among the relative prices to be averaged, the median may give more significant results than the arithmetic mean. (4) When the number of commodities included in the index number is small, however, medians may occasionally prove erratic, representing less the general trend of prices than the peculiarities of the data from which they are made. (5) If the index number is designed for the public at large, the familiarity of arithmetic means is

an argument in their favor; but it counts for nothing in the case of figures intended for specialists. (6) Often the usefulness of a new index number may be enhanced without detriment to its special purpose by throwing it into a form directly comparable with that of index numbers already in existence. Then, of course, not only the form of average but also the base period employed in making the existing series has special claims for imitation. (7) Finally, the desirability of making index numbers that can be shifted from one base to another without raising difficult problems of interpretation deserves more consideration than is commonly accorded it. On this count the score is in favor of the geometric mean. If geometric means were invariably used, all index numbers could readily be compared with one another, whatever the bases on which they were originally computed. And that would be a great gain to all students of prices.[1]

Coupled with this summary, the student should bear in mind the advantages of the aggregative.

135. The Ideal Formula.—Irving Fisher maintains that a good index number should conform to two tests: (1) the time-reversal test, (2) the factor-reversal test.

The time-reversal test is one in which "the index number calculated forward should be the reciprocal of the index number calculated backward."[2]

The arithmetic mean does not conform to this test, as indicated in Table LXVII, in which the relatives on both bases indicate an increase. The geometric mean, however, for 1924 on the 1913 base (70.7) is the reciprocal $\left(\frac{100}{70.7}\right)$ of the index for 1913 on the 1924 base, which equals 141.4.

The factor-reversal test is analogous to the time-reversal test. Just as our formula should permit the interchange of the two times without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent results—*i.e.*, the two results multiplied together should give the true value ratio.[3]

Professor Fisher found that, while the geometric mean conformed to the time-reversal test, it did not conform to the factor-reversal test, but had considerable bias in weighting unless rectified by cross-weighting.

[1] U. S. Bureau of Labor Statistics, *Bull.* 284.

[2] Fisher, Irving, "The Making of Index Numbers," p. 70.

[3] *Ibid.*, p. 72.

In order, therefore, to rectify the bias in weighting, the following formula is given by Professor Fisher as being an "ideal formula" for accuracy in computation.

$$P = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

in which p_1 = the price of any commodity in any given year or period.

p_0 = the price of that commodity in the base year.

q_1 = the quantity of that commodity in the given year.

q_0 = the quantity of that commodity in the base year.

The above formula indicates the weights to change from year to year so that the resulting index may be a result of changes in weights and also changes in the data themselves. It would, therefore, appear that weights and averages are complementary to one another and one should not speak of the correct type of average to use unless it is linked up with the correct weighting system or, as Professor Fisher states:

> Hitherto writers have been debating the "best type" (whether arithmetic, geometric, or median) by itself, the "best weighting" by itself, and the bearing on these of the distribution of price relatives. But from our study it should be clear that it makes little difference what type we start with, or what the weighting is (so long as it is systematic), or what the distribution of price relatives may be so long as we "*rectify*" the formulae and so eliminate all these sources of distortion or one sidedness.[1]

Criticism has been raised against this formula, which has been developed independently by three statisticians. The criticism refers to the term "ideal," which Professor Fisher claims for the formula. All statisticians agree that it is the fitting formula to use when measuring the equation of exchange. Objections have been made to the changing weights and the complexity of computation in contrast to the geometric mean, which has constant weights.

The criticisms merely emphasize the fact that index numbers must be constructed around a definite purpose with tools which are fitting to the purpose itself.

IV. SUMMARY

A concise summary of the preceding chapters on index numbers is almost impossible, since the subject is so controversial, but

[1] *Ibid.*, p. 364.

it may here suffice to state briefly the principles involved in the construction of an index number.

An index number may be expressed as a relative either to a fixed-base or a chain system, and as an aggregative. It is a number which characterizes a group.

The constituent parts and samples should be such as are fitting to the purpose of the study. Their relative importance should give rise to a system of weighting which must be logical.

The analysis of the data may require the use of averages, and since averages have different attributes only those should be chosen which are in accordance with the purpose of the study and do not vitiate the system of weighting.

Indices are important instruments in measuring group changes, and therefore should be carefully made and interpreted.

References

Barnet, G. E.: "Index Numbers of Total Cost of Living," *Quar. Jour. Econ.*, February, 1921.

Bowley, A. L.: "Elements of Statistics," chap. IX.

Davenport, C. B.: "Best Index to Build," *Quar. Pub. Amer. Stat. Assoc.*, September, 1920.

Davies, G. R.: "Introduction to Economic Statistics," chaps. III and IV.

Day, E. E.: "An Index of Physical Volume of Production," *Rev. Econ. Stat.* September, 1920, and January, 1921.

Douglas, P.: "The Wage Studies of the National Industrial Conference Board," *Quar. Pub. Amer. Stat. Assoc.*, September, 1921.

—— and Lamberson, E.: "The Movement of Real Wages," *Amer. Econ. Rev.*, September, 1921.

Fisher, Irving: "Making of Index Numbers."

——: "Purchasing Power of Money," pp. 198–234, 385–430.

Kelley, Truman L.: " Certain Properties of Index Numbers," *Quar. Pub. Amer. Stat. Assoc.*, September, 1921.

——: "Statistical Method," chap. XIII.

King, W. I.: "Elements of Statistical Method," pp. 178–185.

Knibbs, G. H.: "Price Indexes, Their Nature and Limitations, the Technique of Computing Them and Their Application in Ascertaining the Purchasing Power of Money," Commonwealth Bureau of Census and Statistics," Labour and Industrial Branch, *Report* 9, 1918.

Macaulay, F. R.: "Making and Using Index Numbers," *Amer. Econ. Rev.*, March, 1916.

Meeker, Royal: "On Best Form of Index Numbers," *Quar. Pub. Amer. Stat. Assoc.*, September, 1921.

Mitchell, Wesley C.: "Index Numbers of Wholesale Prices in the United States and Foreign Countries," U. S. Bureau of Labor Statistics, *Bull.* 284.

Persons, Warren M.: "Fisher's Formula for Index Numbers," *Rev. Econ. Stat.*, May, 1921.

SECRIST, H.: "An Introduction to Statistical Methods," chap. IX and X.
——: "Readings and Problems in Statistical Methods," chap. VII.
SNIDER, J.: "Wholesale Prices in the United States, 1866-1891," *Rev. Econ. Stat.*, April, 1924.
SNYDER, CARL: "A New Index of Volume of Trade," *Quar. Pub. Amer. Stat. Assoc.*, December, 1923.
YOUNG, ALLYN A.: "The Measurement of Changes of the General Price Level," *Quar. Jour. Econ.*, August, 1921.
U. S. Bureau of Labor Statistics, *Bulls.* 156, 181, 320, 334, and 335.

Problems

1. Differentiate between an index number and a price relative.

2. The average cost in the United States (1913) for specified articles of food, together with their weights for a family of five, are given below:

Item	Weight per month	Average price in United States for year 1913, cents
Sirloin steak	8 pounds	25.4 per pound
Rib roast (beef)	12 pounds	19.8 per pound
Lamb (leg)	4 pounds	18.9 per pound
Pork chops	4 pounds	21.0 per pound
Bacon (sliced)	4 pounds	27.0 per pound
Lard	8 pounds	15.8 per pound
Eggs (fresh)	3 dozens	34.5 per dozen
Milk	50 quarts	8.9 per quart
Butter	10 pounds	38.3 per pound
Bread	62 pounds	5.6 per pound
Flour	30 pounds	3.3 per pound
Beans (navy)	8 pounds	10.5 per pound
Sugar	20 pounds	5.5 per pound
Potatoes	70 pounds	1.7 per pound
Coffee	3 pounds	29.8 per pound
Tea	1 pound	54.4 per pound

Construct an index which measures the monthly changes in the cost of the food for the years 1914 to 1924 (*Source*: "*Retail Prices of Food*" in *Monthly Labor Review issued by Bureau of Labor Statistics*) by means of

a. Relatives to the fixed base 1913.
b. Chain relatives.
c. Continuous chain to 1913 base.
d. Aggregatives.
e. Aggregatives reduced to relatives on 1913 base.
f. Aggregatives reduced to relatives on 1924 base.

g. Relatives in (*a*) expressed as percentage of the mean of the relatives for the monthly relatives for 1924.

3. Why do (*f*) and (*g*) above differ even though they represent the same data reduced to the same base?

4. Why, in the above problem, are retail prices used in preference to wholesale?

5. What are the advantages of the aggregative index?

6. Why was weighting necessary in the above problem, and how, approximately, were the weights derived? (For weights in minimum budget consult *Monthly Labor Review,* June, 1920.)

7. What would be your criticism if the above index were used to measure cost of living? Explain.

8. What are the disadvantages of

a. A fixed-base index.

b. Chain-relative index.

9. How does the aggregative remedy the defects in problem 8?

10. What average or averages, if any, would you use in the construction of an index number? Explain.

11. What are the attributes of the geometric mean in index numbers?

12. What system of weights, if any, would you use in the construction of an index numbers?

CHAPTER XXII

TIME SERIES

136. Introduction.—The methods of comparing data between two periods of time have been discussed. The periods could be of any duration, and, because of this flexibility of the time element, an index number of production for any one industry, if computed as illustrated in the previous chapters, may give erroneous impressions unless it be corrected for certain variations.

137. Type of Fluctuations.—*a. Seasonal.*—The business man in a manufacturing enterprise watches the monthly figure of production carefully. Let it be supposed that instead of using absolute figures, which are usually used, an index is constructed with the January production figures as the base. It may so happen that the index for the current month shows a marked increase over that of the preceding month. Is one justified in assuming that the business is "picking up?" This can readily be answered if it is recalled that various types of industry have their active and their slack months. The current month, therefore, may be the month of greatest activity; the index should show a considerable increase. The index, then, if it is to record business activities for short periods of time, must be corrected for these fluctuations, which are called "seasonal variations." Some series, such as the indices for commodity prices, show little seasonal variation, while others, such as pig-iron production, have marked activity.

b. Secular Trend.—If the period of time be changed from monthly figures to yearly figures, the business man in comparing production for the last fiscal year with that of ten years ago may feel that his business has expanded considerably in the ten years, when measured by the crude index. Business, however, is a relative thing; men measure their efforts by the profits derived in the various industries. Over a period of time capital *tends* to flow into the most profitable fields. The business man, therefore, compares not only the returns in his own field of industry but also industry in general. In ten years industry, in general,

has grown; population has increased, thus increasing the domestic demand for goods, while commerce, in general, has increased. The increase in industry, in general, however, is like the general increase in prices. Some forms of business may register a decline, others just maintain their own, while still others may show all ranges of increase. The question then for the business man is not to find out how much his business has increased in the last ten years, but rather the annual proportional increase in comparison with industry in general, in order that some idea may be given as to the position of the industry from the viewpoint of demand—in other words, to answer the questions: (1) Is the industry maintaining its demand, or is it a passing business? (2) How does its growth compare with that of other industries? To answer these questions the figures must be corrected in such a way as to indicate the growth element or secular trend. Secular trend is long-time change in contrast to short-time change as recorded by seasonal variation. The growth element may be positive, as illustrated by the rapid growth of pig-iron production, or negative, as indicated by the fall in prices for the years 1876 to 1896.

c. Cyclical Variation.—Conditions change with time the relatively short periods of months vary, some being active while others are slack. Industry may grow over a long period of time, but it may encounter temporary setbacks; the yearly figures vary according to the phase of the business cycle. This year may be in the "depression" period, while last year may have been in the "prosperity" phase of the business cycle. The yearly figures, therefore, afford no basis, in themselves, for forecasting future production, unless they are interpreted in terms of cyclical fluctuations.

Statistical control of production requires a production policy which aims to regulate industry, to eliminate, as far as possible, the undue seasonal activities, and particularly to build up a forecaster in order that the phases of the business cycle may be met with the least possible dislocation of business. A business man should know the relation of the cycle of his own industry to that of the general business cycle. Therefore, production figures should be reduced to cyclical fluctuations if cyclic relationships are desired.

138. Summary.—The figures involved in a time series may therefore involve three commensurable variations: seasonal,

secular, and cyclical fluctuations. Production figures should be corrected for the required element; if seasonal figures are required, the secular and cyclical elements should be eliminated. Likewise, if the object is to derive cyclic variations, both the seasonal and secular must be eliminated.

There is a fourth fluctuation which is hardly commensurable in that it is fortuitous—changes arising through abnormal conditions, such as experienced during the Great War.

This introduction merely indicates the problems to be discussed in the chapters which follow, and also emphasizes the fact that business men should not use figures indiscriminately.

CHAPTER XXIII

THE PROBLEM TO WHICH ANALYSIS IS APPLIED

As the raw material upon which to apply the various methods of analysis, the monthly values of the total exports from the United States to Canada for the years 1901 to 1919 are given below in Table LXX. Also included in the table are the yearly totals for the years 1920 to 1923. The monthly figures for these four years were omitted, as their inclusion would make the study too protracted. The nineteen-year period of the study illustrates periods of normal conditions, abnormal, and returning to normal after the War.

In any problem the sources of the raw material should be given as well as a definition of the unit of measurement, also any peculiarities should be noted. Some idea of the type of the report is given below.

Study of Total Exports from the United States to Canada. Monthly Figures for Years 1901 to 1919 inclusive.

The data from which the following study was made were compiled from the *Monthly Summary of Foreign Commerce of the United States* under the section of "Imports and Exports of Domestic and Foreign Merchandise by Countries."

The study is made from the original data, in that the *Monthly Summary* is the official publication, issued by the Bureau of Statistics of the Department of Commerce, from *Custom District Reports*. The figures are exhaustive, with slight modifications for the yearly totals.

The unit of measurement is defined in the original report of 1866, which states "value of our exports always is, or is believed to be, their value here and is expressed in the currency of the United States." In this study the unit is that of $10,000.

The data for the year 1901 to 1903 inclusive were listed under the heading "British North America" and, therefore, from the figures thus given in the official report had to be deducted the exports to Newfoundland; after 1903 the exports were classified under separate headings.

TABLE LXX.—VALUE OF TOTAL EXPORTS FROM THE UNITED STATES TO TO CANADA, MONTHLY FIGURES, 1901–1919 INCLUSIVE

Units of $10,000

Months	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910
January	732	680	775	770	916	1,225	1,362	1,111	1,080	1,538
February	643	726	923	759	803	1,079	1,242	1,070	1,216	1,626
March	773	870	1,089	1,168	1,200	1,313	1,587	1,338	1,440	2,017
April	838	977	1,237	1,107	1,382	1,436	1,781	1,404	1,520	2,191
May	1,080	1,037	1,277	1,353	1,370	1,583	1,804	1,310	1,765	2,076
June	1,031	980	1,229	1,357	1,273	1,495	1,796	1,382	1,678	2,074
July	993	936	1,267	1,280	1,274	1,458	1,654	1,292	1,678	1,991
August	1,127	1,089	1,194	1,237	1,264	1,559	1,680	1,300	1,666	2,295
September	925	991	1,129	1,267	1,267	1,376	1,501	1,263	1,702	2,071
October	1,009	1,008	1,088	1,376	1,293	1,467	1,683	1,334	1,692	2,208
November	819	926	1,051	1,152	1,269	1,503	1,467	1,244	1,768	2,162
December	714	810	906	902	1,180	1,364	1,121	1,215	1,568	1,984
Totals	10,684	11,030	13,165	13,728	14,491	16,858	18,678	15,263	18,773	24,233

Months	1911	1912	1913	1914	1915	1916	1917	1918	1919
January	2,007	2,132	3,273	2,135	1,742	3,830	5,892	5,031	5,571
February	2,077	2,334	3,305	2,135	2,422	3,771	5,384	4,256	4,062
March	2,352	2,964	3,292	2,618	2,255	4,115	7,704	6,952	3,869
April	2,627	3,099	3,742	2,817	2,562	4,900	7,574	7,515	5,941
May	2,627	3,399	3,939	2,958	2,556	4,959	9,098	8,557	5,731
June	2,592	3,370	3,762	2,823	2,956	5,107	9,448	7,698	7,524
July	2,481	3,499	3,422	3,161	2,950	5,009	5,305	6,905	5,529
August	2,706	3,573	3,544	3,349	3,477	5,688	7,139	8,614	6,245
September	2,579	3,373	3,073	2,595	3,190	5,213	5,386	8,932	7,798
October	2,887	3,441	3,310	2,359	3,530	5,535	6,400	8,250	6,798
November	2,660	3,400	3,060	2,352	3,255	6,292	3,616	8,704	7,471
December	2,312	3,061	2,606	1,838	3,604	6,095	5,080	7,272	7,191
Totals	29,907	37,645	40,128	31,140	34,499	60,514	78,026	88,686	73,730

Yearly totals

1920	1921	1922	1923
97,185	59,367	57,686	55,920

The graphic presentation of the yearly totals drawn on logarithmic scale is shown in Chart 55. The logarithmic projection gives a more accurate picture of the proportional increase

LOGARITHMIC CHART OF VALUE OF EXPORTS FROM THE UNITED STATES TO CANADA, 1901–1923

Source: Monthly Summary of Foreign Commerce

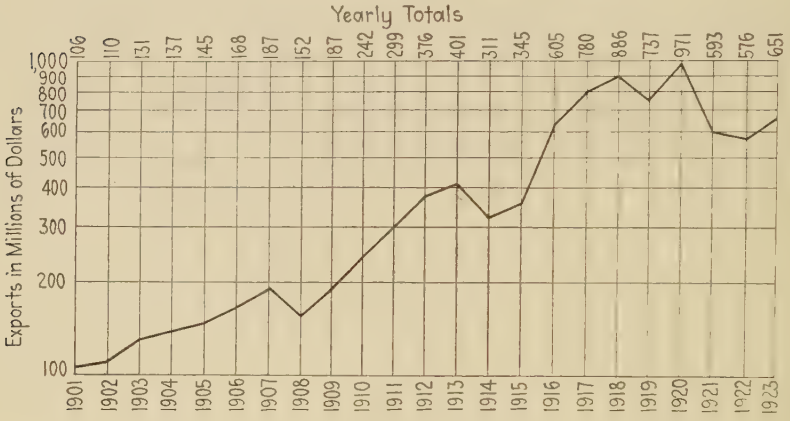


CHART 55.

than the same data plotted on rectilinear paper (see Chart 57). Both charts indicate clearly the cyclic period, 1907 to 1913 for exports.

CHAPTER XXIV

SEASONAL VARIATIONS

139. Introduction.—As already indicated, seasonal variation consists of those fluctuations which occur within a yearly period, and is the type of variation to which the business man gives considerable attention. He is always interested in comparing current production with that of the preceding month. Therefore, to use the figures of the previous month intelligibly he must know the relative degrees of activity of the twelve months of the year.

140. Causes for Monthly Variations.—Seasonal variations are not to be considered as fortuitous, but recurring year after year, not to the same degree, to be sure, but to be found within certain limits. In other words, there is a consistent relationship between monthly figures over a period of time. Chart 58 on page 300 indicates the similarity of monthly variations for the years prior to the war.

The amount of seasonal variation depends upon the type of series—some show marked periods of activity and slump, while others pursue a fairly even course.

Variations may be attributed not only to relative degrees of activity due to seasonal demand over the twelve-month period, but also to the length of the month; the number of non-working days, such as Sundays and holidays; also to change in working conditions, such as the change from the 54- to the 48-hour week. The change in the number of hours worked can be readily corrected in computation, since it only affects the months in the year of change. The number of days worked per month, however, present a more difficult problem in that there is constant change both from one month to the other and also for the same month for different years.

Because of the difficulties presented by the varying lengths of the months, efforts are being made on the part of some—principally Moses Cotsworth—to have an international agreement on a thirty (30) day statistical month, thus giving statistical

months of equal duration. Such a proposal, however, does not obviate the need for the correction of the variance in the number of Sundays in various months.

The discrepancies arising from the difference in the lengths of months could be corrected by obtaining daily averages and working with these, but the process is too time-consuming and for general purpose is not worth the effort. It must be remembered, at all times, that seasonal indices are approximations and, therefore, the errors arising through this inequality would, if corrected, not give an infallible index. Furthermore, if a long period of time is used in deriving the seasonal indices, some of these changes, such as the number of Sundays, will tend to counteract one another.

The month, whether it has twenty-eight or thirty-one days with four or five Sundays, will be the period used in this chapter for the measurement of seasonal variations. The methods used will, of course, be applicable to any study involving monthly or quarterly figures.

With an understanding of the causes for monthly variations in production, and the limitations which these causes place upon the use of the seasonal indices as a rigid guide for production, it is necessary to discuss the methods involved in deriving seasonal indices.

141. Methods in General.—Since time series may involve short- and long-time fluctuations together with cyclical fluctuations, the measurement of seasonal variation must devise some process whereby the secular and cyclical fluctuations are eliminated from production or export figures.

Many business men realize that, owing to seasonal activity, it would be foolish to compare the production figures of the past month with those of the preceding month in order to see if business were "picking up." They therefore compare production with the same month of the preceding year. This does afford a better comparison, but the preceding year may have been in the period of "depression," so that the monthly figures might contain in themselves an element of the cyclic fluctuations. Such a method, therefore, does not afford a strict basis for comparison.

Other business men, in order to get a more comparable basis, compare the production of the past month with the average production of the same month over a period of years.

By using the ten-year monthly averages, seasonal fluctuations are automatically allowed for, and by always taking the ten preceding years as the base provision is made for normal growth.[1]

Such a method does tend to eliminate the cyclic fluctuations because the mean is usually obtained from both eras of prosperity and depression. By consulting Chart 55 it will be noticed that in a ten-year period there has been a decided growth element. The average however, for a ten-year period may not entirely allow for the growth element. While this method is an improvement over the above, it hardly conforms to rigid statistical practice.

The discrepancies that might arise through failure to recognize the secular and cyclic elements in time series may be cited by the criticism against certain results as discussed by Stuart Chapin:[2]

> Annual or seasonal fluctuations were computed from twelve of these (dependency) series (for the years, 1914 to 1916) by Rubinow, who used a method of simple averaging. No allowance was made for the effect of secular trend upon the seasonal variation in the series that were complicated by a definite secular trend and a violent cyclical disturbance. It is the object of this paper to show the practical effects of recomputing the seasonal variation by allowing for the influence of secular trend and of the business cycle . . .
>
> The effect of our correction, therefore, has been to push the peak load into the winter months; that is, the high point of demand on relief agencies falls in November and December rather than in December, January, and February as it would appear from Rubinow's series.

From the above quotation it becomes evident that some method must be used which will eliminate both the secular and cyclical variations. Two such methods are available:

1. The method of monthly means.[3]
2. The link relative method.[4]

The method used in this study will be that of the link relative as developed by Warren M. Persons. For an illustration of the first method the student is referred to G. R. Davies, "Economic Statistics," pages 116–120.

[1] COPELAND, M. T., "Business Statistics," p. 126.

[2] "Seasonal Variation in Dependency," *Amer. Stat. Pub.*, pp. 911–914, September, 1923.

[3] "The Method of Monthly Means for Determination of Seasonal Variation," *Amer. Stat. Pub.*, p. 341–349, September, 1922.

[4] PERSONS, WARREN M.: "*Rev. Econ. Stat.*," vol. I, pp. 18–31.

142. Link Relative Method.—*a. Introduction.*—Since the problem is to find a measure which will show monthly relationships and yet eliminate cyclical variations and secular trend, it is evident that the relationship of one month to another, December to November, etc., should form the basis of computation, and that a sufficient number of these monthly relationships should be taken in order that their average will be one derived from all phases of the business cycle, thus eliminating cyclical variations. The problem of the elimination of secular trend will be discussed later, suffice it here to state that the growth element in monthly periods is very small. The correction required is usually very small.

b. Monthly Relatives.—The first step is suggested by the title "link relative." The figures for each month are *divided* by those of the *preceding month,* which results in monthly relatives, the absolute numbers being lost.

The application of this method to the data of Table LXX, (page 281) gives a relative for January, 1901

$$\left(\frac{\text{January 1901}}{\text{December 1900}} = \frac{732}{806} = 90.8\right),$$

or 91, since the relatives are recorded to the nearest 1 per cent. The February relative for the same year $= \frac{643}{732} = 88$:

$$\text{January 1903} = \frac{\text{January 1903}}{\text{December 1902}} = \frac{775}{810} = 96.$$

This process is continued for the entire period of the study, the tabulated results being given in Table LXXI.

c. The Representative Relative.—Table LXXI presents the link relatives of the data of each month expressed as a relative of the data of the preceding month for a period of nineteen years. In reading down any one column of the relatives for any one month for nineteen years, variations in the relatives themselves will be observed; these generally indicate changed business conditions. Some years are good, while others (1908 and 1913) are relatively poor ones, as shown by the fact that the seasonal variation between different months for these years is not so marked, as evidenced by reading the rows for the years 1907 and 1913. The relatives, therefore, measure business under all conditions, and the question arises: Which relative of the nineteen may be taken as representative of each month? Of all

TABLE LXXI.—LINK RELATIVES OF EXPORTS FROM THE UNITED STATES INTO CANADA, 1901–1919

For computation of the seasonal variations

Year	Jan. / Dec.	Feb. / Jan.	Mar. / Feb.	Apr. / Mar.	May / Apr.	June / May	July / June	Aug. / July	Sept. / Aug.	Oct. / Sept.	Nov. / Oct.	Dec. / Nov.	Jan. / Dec.
1901	91	88	120	121	129	96	96	114	82	109	81	87	
1902	95	107	120	112	106	95	95	116	90	102	92	87	
1903	96	119	108	114	104	96	93	94	94	96	97	86	
1904	85	98	154	94	121	100	84	96	102	108	84	86	
1905	102	88	148	115	99	93	100	99	100	102	98	93	
1906	104	88	121	109	110	96	97	106	88	106	102	91	
1907	100	90	136	112	101	94	92	101	89	112	87	76	
1908	99	96	125	105	93	105	93	100	97	105	93	98	
1909	89	112	118	105	126	95	100	99	102	99	104	89	
1910	98	106	124	108	95	100	95	115	90	106	97	92	
1911	101	100	113	112	100	98	109	95	95	111	92	87	
1912	92	109	127	104	109	99	104	108	94	102	99	90	
1913	102	109	99	114	105	95	91	104	87	108	92	85	
1914	82	100	123	107	105	95	110	107	77	92	99	78	
1915	96	139	93	114	100	115	99	118	91	119	92	117	
1916	106	98	100	119	101	123	98	115	91	106	113	97	
1917	97	91	143	98	120	104	55	134	78	121	56	140	
1918	92	84	163	167	114	90	89	125	104	92	106	84	
1919	74	73	95	153	96	128	73	113	124	87	110	96	
a. Medians	96	98	121	112	105	96	96	107	91	107	97	89	
b. To January base	100	98	118.6	132.8	139.4	133.8	128.5	137.5	125.1	133.9	129.7	115.6	111.0
c. Correction	...	0.9	1.8	2.7	3.6	4.5	5.5	6.4	7.4	8.3	9.2	10.1	11
d. Corrected indices to January	100	97.1	116.8	130.1	135.8	129.3	123.0	131.1	117.7	125.6	120.5	105.5	100

Arithmetic mean of corrected relatives to January base $= \frac{1,432.5}{12} = 119.4$, which gives the mean monthly average for the year, which figure, when divided into the relatives in row *d*, gives the final seasonal indices, *e.g.*, January index is $\frac{100}{119.4} = 83.7$. February $\frac{97.1}{119.4} = 81.3$, etc.

FINAL MONTHLY INDICES

Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
83.7	81.3	97.8	108.9	113.7	108.3	103	109.8	98.9	105.1	100.8	88.3

the nineteen January relatives, which one represents January under strictly normal conditions? Naturally, it would not be the relative representing the extreme item. Since this is the result of abnormal or cyclic conditions. The determination of the relative to use is based largely upon the properties of the various averages.

In that extreme items are not representative, the arithmetic mean average of the relatives should not be used, since the mean gives weight to extreme items. The geometric mean also emphasizes the abnormal. The mode, being the measurement of the item of greatest frequency, would appear to be the average to use, but because series may be bimodal or have no definite point of concentration and the mode requires, to be accurate, considerable data, this average is not used. The median, on the other hand, gives no weight to extremes, but is the mean between extremes. This seems to be the logical average to use in finding the representative monthly relative.

The medians of the respective link relatives are shown in row *a* of Table LXXI. The method of derivation of the median is familiar to the reader. The nineteen relatives for any one month are arranged in ascending order so that the tenth relative is the median. The January relative was obtained in the following way. The relatives for January for all nineteen years were arranged in ascending order, the tenth item being 96.

Relatives	Relatives
74	96 Median
82	97
85	98
89	99
91	100
92	101
92	102
95	102
96	104
	106

If the relatives are numerous and their range considerable it is better to construct a table with class intervals and tabulate the relatives into the table and so derive the median by the method used for frequency tables.

d. Relative to January Base.—With the selection of the medians the problem of seasonal variation has not yet been solved, since the relationships are, as yet, only those of one

month to the preceding one, whereas it is desirous to know how December relates itself not only to November but also to the remaining ten months of the year. This necessitates the making of a continuous series of the medians; to express them as percentages of some fixed-base period. The first step is to assume January to be the fixed base with a value of 100.

Since the median for each month is devised by reference to the preceding month $\frac{\text{January}}{\text{December}}$, the medians for each of the eleven months should not be expressed *directly* as a percentage of January but rather in its relationship to the relative of the preceding month which has already been reduced to the January base. The median for February from the link relatives was found to be 98. Therefore, if January is given a value of 100 the new February relative equals $98 \times 100 = 98$. Since March has a median relative of 121, and February has a relative of 98 when it is measured from the January base, March must have a relative of $100 \times 89 \times 121$ (January base $\times$ February $\times$ March medians), or 117.6 when referred to the January base; in other words, the medians are multiplied continuously from January. The April relative to January base equals January base $\times$ February $\times$ March $\times$ April medians $= 132.8$.

The relatives to the January base are shown in row *b*, Table LXXI. Naturally, if the April relative is derived by multiplying January base $\times$ February $\times$ March $\times$ April medians, thus giving a new relative of 132.8, the May relative could be derived by multiplying the *old* May relative 105, row *a*, by the *new* April relative 132.8, row *b*, because the relatives of the previous months were used in the computation of the new April relative. Row *b*, Table LXXI, was derived by multiplying the old relative for any one month by the new relative for the preceding month. This is shown by $\times$ in the table which links the numbers used in deriving the new index.

e. Correction for Secular Trend.—The relatives to a January base should be such that if the December relative to the January base be multiplied by the January median, the result should be 100, the value which has been assigned to January of any year. In the problem, December has a value of 115.6 (row *b*) while the January median is 96; the product of these two yield a result of 111, which means that in a period of twelve months, from January to the next January, the base has increased 11 per

cent. This increase may be attributed to secular trend, the growth element showing itself in the original link relatives and thus affecting the medians. The growth element is positive. Therefore, the more recent figures, December in contrast to January of any given year, tend to have larger figures than would have been the case if there had been no growth. Hence the relatives for the later months of the year contain a greater element of secular trend. Conversely, if the trend had been negative the earlier months of the year would have been larger than the later months, so that the medians for the first few months of the year would have been larger than if normal conditions prevailed.

Seasonal variations should be devoid of both cyclic and secular trends. Therefore the increase of 11 per cent should be deducted in some way from the relatives so as to afford an even distribution of the excess. The easiest way is to deduct one-twelfth of the excess from February relative to the January base, two-twelfths from March—and twelve-twelfths from the second January index, thus reducing it to 100. This correction is shown in row *c*, Table LXXI while row *d* gives the corrected indices.

f. Indices from Mean of Corrected Indices.—The final step in the computation of the seasonal indices is that of obtaining the mean of the corrected monthly indices. The mean of these figures is taken because it is desired to know how January figures conform to the yearly average. The mean of the corrected indices is 119.4, which sum, when divided *into* each of the corrected indices, as shown in Table LXXI, indicates the final monthly seasonal variations.

g. Logarithmic Derivation.—For those familiar with logarithms Table LXXII presents an easy method of computing the seasonal indices, once the medians have been determined. This method is also more accurate in that the correction for secular trend should be such that it follows the compound-interest curve; that is, the secular trend item is cumulative so that it tends to be smaller in the early months and larger in the latter. The secular trend correction by the logarithmic method is the log of 0.0453, as shown in column D. This is distributed as shown in the same column; one-twelfth of log 0.0453 is deducted from the log of the February relative to the January base, two-twelfths

TABLE LXXII.—COMPUTATION OF SEASONAL VARIATION BY MEANS OF LOGARITHMS

	Median relatives	Log of medians	Logs to January base	Compound interest correction	Corrected logs to January base	Antilog to January base	
	A	B	C	D	E	F	
January	96	1.9823	2.0000	0.0000	2.0000	100	
February	98	1.9912	1.9912	0.0038	1.9874	97.1	
March	121	2.0828	2.0740	0.0076	2.0664	116.7	
April	112	2.0492	2.1232	0.0113	2.1119	129.5	
May	105	2.0212	2.1444	0.0151	2.1293	134.6	
June	96	1.9823	2.1267	0.0189	2.1078	128.2	
July	96	1.9823	2.1090	0.0226	2.0864	122.0	
August	107	2.0294	2.1384	0.0264	2.1120	129.4	
September	91	1.9590	2.0974	0.0302	2.0672	116.8	
October	107	2.0294	2.1628	0.0340	2.0927	123.9	
November	97	1.9868	2.1136	0.0378	2.0758	119.0	
December	89	1.9494	2.0630	0.0416	2.0214	105.1	1,422.3
January	96	1.9823	2.0453	0.0453	2.0000	100	

The arithmetic mean of antilogs $= \frac{1,422.3}{12} = 118.5$, which is the mean monthly average for the year, which figure when divided into the respective corrected *relatives* to the January base, column F, gives the final, seasonal indices, *i.e.*, $\frac{\text{January}}{118.5} = \frac{100}{118.5} = 84.3$; $\frac{\text{February}}{118.5} = \frac{97.1}{118.5} = 81.9$, etc.

FINAL MONTHLY INDICES

Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
84.3	81.9	98.5	109.2	113.6	108.1	102.9	109.2	98.5	104.5	100.4	87.7

from March, etc. The table itself indicates all the steps in computation.

h. Graphic Presentation.—The final indices are shown graphically in Chart 56, which portrays both the indices corrected by the arithmetic and those by the logarithmic. The two curves illustrate the discrepancy of the arithmetic method which gives lower values for the earlier months and higher for the later months when compared to those given by the logarithmic computation. This error, as already indicated, is due to the secular trend being cumulative. A dollar invested at 4 per cent compound interest shows smaller accumulations the first few years

than a dollar invested at a straight 5 per cent rate. However, in time, the interest accumulates so that the return on the compound-interest investment is greater than that on the straight 5 per cent rate. This illustration is analogous to the secular trend. The discrepancy, however, is negligible, so that the arithmetic method will suffice for those not familiar with logarithms.

SEASONAL FLUCTUATIONS OF EXPORTS FROM UNITED STATES TO CANADA
Based upon Tables LXXI and LXXII

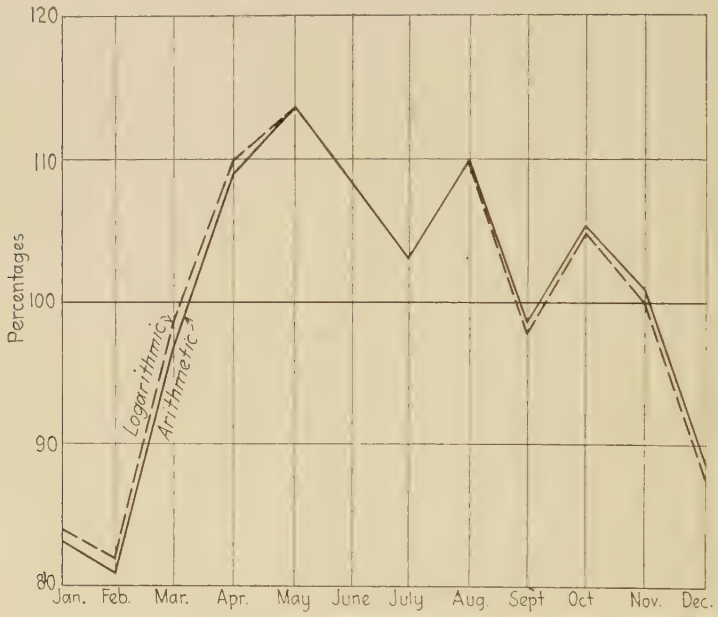


CHART 56.

143. Summary.—The derivation of seasonal indices may seem somewhat complicated and it is rather a complex process, but the steps in themselves are comparatively easy though time-consuming. They are, likewise, logical in their development and there should be little difficulty in using them if it is remembered that cyclical and secular trend must be eliminated in seasonal indices. The method may be summarized as follows:

1. Express the figures for each month as a percentage of the previous one $\frac{\text{February}}{\text{January}}$.

2. Obtain the medians of the relatives for the respective months.

3. Express each median, by multiplying serially, as a percentage of the fixed January base.

4. Correct the percentages from (3) for secular trend as shown by the increase or decrease between January relatives.

5. Obtain the mean of the corrected indices.

6. Express each of the corrected indices as a percentage of the mean (5).

References

COPELAND, M. T.: "Business Statistics."

CRUM, W. L.: "Use of the Median in Determining Seasonal Variation," *Quar. Pub. Amer. Stat. Assoc.*, March, 1923.

DAVIES, G. R.: "Introduction to Economic Statistics," pp. 116–123.

HART, WILLIAM L.: "The Method of Monthly Means for Determination of Seasonal Variation," *Quar. Pub. Amer. Stat. Assoc.*, September, 1922.

JORDAN, D. F.: "Business Forecasting," p. 90.

KEMMERER, S. W.: "Seasonal Variations in the Relative Demand for Money and Capital in the United States," *Report Nat. Monetary Comm.* 1910.

KING, W. I.: "Elements of Statistical Method," pp. 192–193. *Rev. Econ. Stat.*, vol. I, pp. 18–31.

Problems

1. What is seasonal variation?

2. Why is it important for the business man to know the relative monthly activity of his business?

3. What other fluctuations should be eliminated from data before seasonal indices are derived?

4. How does the link relative in the link relative method tend to eliminate secular trend?

5. How does the link relative method, however, finally eliminate all secular trend?

6. Why is the median used in determining the representative monthly link relative?

7. From the *Monthly Labor Review* of the United States Bureau of Labor Statistics obtain the monthly retail price, for the United States, of butter for the years 1913 to 1924.

8. By the method of link relatives derive the seasonal indices for the data of problem 7.

9. Illustrate by graphic methods the result of problem 8.

10. Why may the average of the production for the same month of the preceding ten years prove a defective standard from which to measure current production?

11. *a.* Record the monthly and yearly totals of the number of business failures in the United States for the period 1913 to 1924. (*Source: Bradstreet's or Dun's.*)

b. Compute the seasonal indices for business failures.

c. Illustrate (*b*) graphically.

d. At what months in the year are business failures, on the average, at their lowest and the highest?

CHAPTER XXV

SECULAR TREND

144. Introduction.—Industries, like individuals, have their youth, maturity, and old age. Some industries, like individuals, seem to possess such vitality that old age creeps upon them very slowly, others, on the other hand, become prematurely aged,

YEARLY VALUE OF EXPORTS FROM THE UNITED STATES TO CANADA, 1901–1919

Unit $1,000,000

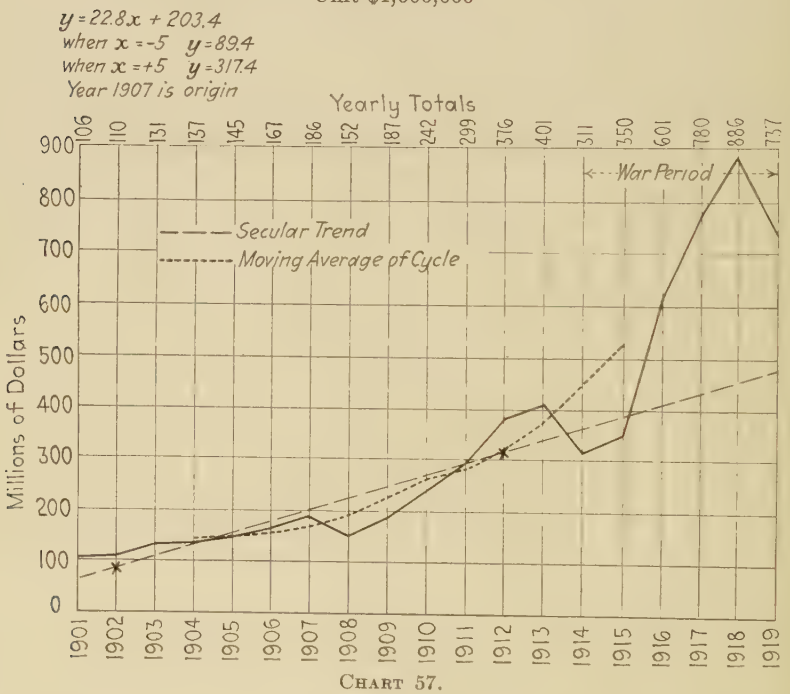


CHART 57.

while still others show rapid growth in their youth only to be cut off before maturity.

The business man is interested not only in his seasonal production figures, which after all are short-time variations, but also in the long-time change. He should try to answer the following

questions: (1) Is the demand for my product maintaining itself as the years pass by; or is the demand increasing? (2) Is the demand, perchance, falling off gradually, or has it declined rapidly? These are vital questions to the business man and an affirmative answer to the second question should be a signal for the planning of the withdrawal of capital from the industry. The answers to these questions involve the measurement of secular trend or normal growth, the rate at which the business is growing year by year.

Secular trend, as the questions above imply, may be either positive or negative, positive when the industry or trade continues to grow. This is indicated by the monthly figures of the problem in hand as plotted on Chart 58. The trend is more apparent when the yearly totals are used as in Chart 57.

The term "normal growth" was used in explaining the term "secular trend" and should not be confused with mere growth. Growth may arise through abnormal conditions, such as exhibited by the rapid rise of the curve in Chart 57 for the years 1915 to 1918. This growth was due to war conditions. Such conditions were hardly comparable to those existing prior to 1914, and therefore are not included in the derivation of secular trend. This statement should also make it clear that the measurement of secular trend involves the measurement of normal growth produced under practically homogeneous conditions; that is, *the measurement should only be derived from those years during which there is a considerable degree of homogeneity of conditions.*

145. Secular Trend as a Forecaster.—Secular trend, therefore, is really the measurement of the growth in past production, when production figures have been corrected for the seasonal and cyclic elements. It measures the yearly production which would have occurred if the growth between the two given periods of time were distributed evenly over the intervening years. Graphically, secular trend is commonly represented by a straight line which is called the line of secular trend. Since the line is computed from past production figures and relates to conditions which are more or less homogeneous, it becomes a very dangerous tool if projected indiscriminately into the future in order to forecast future growth. However, data that are a direct function of population may warrant some measure of forecasting demand, for, as Raymond Prescott, in a study of various industries and population growth, stated:

These results seemed to indicate that there is a general law of growth that would fit the growth of population, and also some products, perhaps all that are a direct function of the growth of population.[1]

The automobile industry has had a remarkable growth and there are some people who are so optimistic as to state that the rate of increase in the production of automobiles will maintain itself for some years. Such a statement implies, if the secular trend were determined for the industry, that a given number would be considered as the yearly addition to past production. By the use of this figure, knowing present production, the production of cars for 1930 could be forecasted. Invention, however, which influences demand, may cause premature old age to seize the apparently young and virile industry. A new invention may create a new demand and, at the same time, cause the withdrawal of demand from an old-established industry. The automobile "killed" the bicycle industry, so who knows if the automobile will be the chief means of locomotion in 1930. Present demand, therefore, does not justify the forecasting of a continuance of patronage.

146. The Measurement of Secular Trend.—From the preceding paragraphs, it is evident, only those years should be used in the computation of secular trend for which there is a marked degree of homogeneity of conditions. In connection with the problem used as an illustration in these chapters, the conditions prevailing from 1901 to 1913 were similar, whereas from 1914 to 1918 abnormal conditions prevailed. Therefore, the years 1901 to 1913 inclusive, form the basis of computation.

Likewise, the seasonal and cyclic fluctuations are to be eliminated. In order to eliminate the seasonal element the evaluation of secular trend should be derived originally from the *yearly totals*, because the yearly figures cover up the seasonal element. The cyclical element may be eliminated by using at least the period of the cycle, so that all phases of the cycle are included in the period used as the basis for computation. The period used should preferably be longer than that of the cycle, if homogeneity of conditions prevail. The cycle for the exports from the United States to Canada is found to be of seven year's duration as shown by Chart 57, the crests and troughs falling on the years 1901 to 1913 and 1908 to 1914, respectively. Since war condi-

[1] PRESCOTT, RAYMOND B., "The Law of Growth in Forecasting Demand," *Amer. Stat. Pub.*, p. 471, December, 1922.

tions prevailed in the latter part of 1914 the period of 1907 to 1913 was chosen to represent the cycle which was of seven years' duration. Furthermore, since normal conditions prevailed for the years 1901 to 1913, these thirteen years are used as the basis for the derivation of secular trend.

In any problem the cycle may be determined by plotting monthly or yearly totals, preferably yearly totals, and measuring the time elapsed between the major crests or troughs as exhibited in the graph. If the study covers a considerable period of time there may be several cycles of unequal duration. The average cycle, provided it is representative of the *more recent data*, should be used.

The data may now be analyzed by any one of the following methods.

1. Graphic method—Freehand.
2. Method of moving averages.
3. Line of least squares.
4. Method of moments.
5. Parabola trend and compound-interest curve.

In this text only the first three methods are touched upon because the fourth method yields results identical with the third. Method No. 5 is used when the secular trend is not linear but takes a curvilinear conformation. For most studies the straight-line fit suffices.

147. Graphic Methods.—The freehand method is a graphic method by means of which a straight line is drawn through the data not by mathematical determination, but by inspection, so that the sum of squares of the deviations of the actual data from the line is at a minimum. The line so drawn is the line of secular trend. This method, however, is but an approximation and should be used sparingly.

148. Secular Trend by Moving Average.—The principle of the moving average was discussed in connection with the mode (Part I). The same principle is applied in the derivation of the secular trend. The only qualification is that the period used should be that of the business cycle, in order that the mean will be one that will allow for both prosperity and depression, and so tend to eliminate the cyclic fluctuations.

Because cycles are not always of equal duration, the application of any one cycle to a study involving a long period of time may not eliminate the cyclic element.

In the problem the cycle was found to be of seven years' duration. Therefore, by obtaining the mean of any seven-year period it is assumed that the total from which the mean is computed was derived from all phases of the business cycle. Such a mean, therefore, gives as much weight to the depression period as that of prosperity. Commencing with the first seven years of the data, a mean 141 $\left(\frac{986}{7}\right)$ was obtained, which was recorded as the value for the year 1904, which is the midpoint of the seven-year period which yields the total of 986. To derive the secular trend figure for 1905 the total for 1901 (106) is omitted from the seven-year total, while that for 1908 (153) is added. If the difference (153 to 106), 47, is added to the previous seven-year total (986), thus giving a new total

TABLE LXXIII.—SECULAR TREND BY MOVING AVERAGES OF SEVEN-YEAR CYCLE APPLIED TO EXPORTS FROM THE UNITED STATES TO CANADA, 1901–1919

Units of $1,000,000

Year	Value	Average of seven years
1901	106	
1902	110	
1903	132	
1904	137	141
1905	145	148
1906	169	159
1907	187	174
1908	153	197
1909	188	231
1910	242	264
1911	299	281
1912	376	309
1913	401	368
1914	311	445
1915	345	529
1916	601	531
1917	780	
1918	887	
1919	737	

of 1,033, the 1905 moving average may be obtained by dividing this sum by 7. This process is continued for all years of the study, as shown in Table LXXIII.

The method of moving averages, as illustrated, does not permit the derivation of secular trend for the extreme years of the study (see Table LXXIII). In that the more recent years are usually more important than the more remote, it would be better if the trend could be derived for the more recent years, even at the expense of computations for the more remote. Such a method is given in the *Harvard Review of Economic Statistics*, Vol. I pages 10 to 12. This method, however, necessitates the use of relative figures.

The moving average requires considerable time in computation and does not yield as accurate results as the method of least squares. Therefore it is not generally used in the derivation of secular trend.

149. Method of Least Squares.—The line of least squares is that line which, when drawn on the graphic presentation of the data, reduces the sum of the squares of the actual deviations of the data from that line to a minimum. It is a mathematical determination in contrast to the free-hand method.

The method of computation is familiar to the student, since it was discussed in connection with the coefficient of correlation and the lines of regression (Part I).

The line of secular trend, if it be linear, will, when plotted, incline to the base at a definite angle. The inclination will be that of 45 degrees if the association between x (years) and y (items) is perfect, provided the scales used in the graphic presentation are in proportion to their respective standard deviations, so that $\frac{\Sigma xy}{\Sigma x^2} = 1$.

It is not the degree of association or correlation that is required for secular trend, for correlation is assumed to be perfect, annual items increasing always by the same increment. What is required is the "ratio of variation" (see Line of Regression Part II, Chap. XVIII) in order that the amount of annual normal increase and also line of secular trend may be derived.

The ratio of variation may be derived from the equation

$$\text{Slope or tangent} = \frac{\Sigma xy}{\Sigma x^2}.$$

MONTHLY EXPORTS (VALUES) FROM THE UNITED STATES TO CANADA, 1901–1919, MONTHLY DATA
Units of $1,000,000
Line of Secular Trend $y = .16x + 16.9$, Data in Table LXXVI

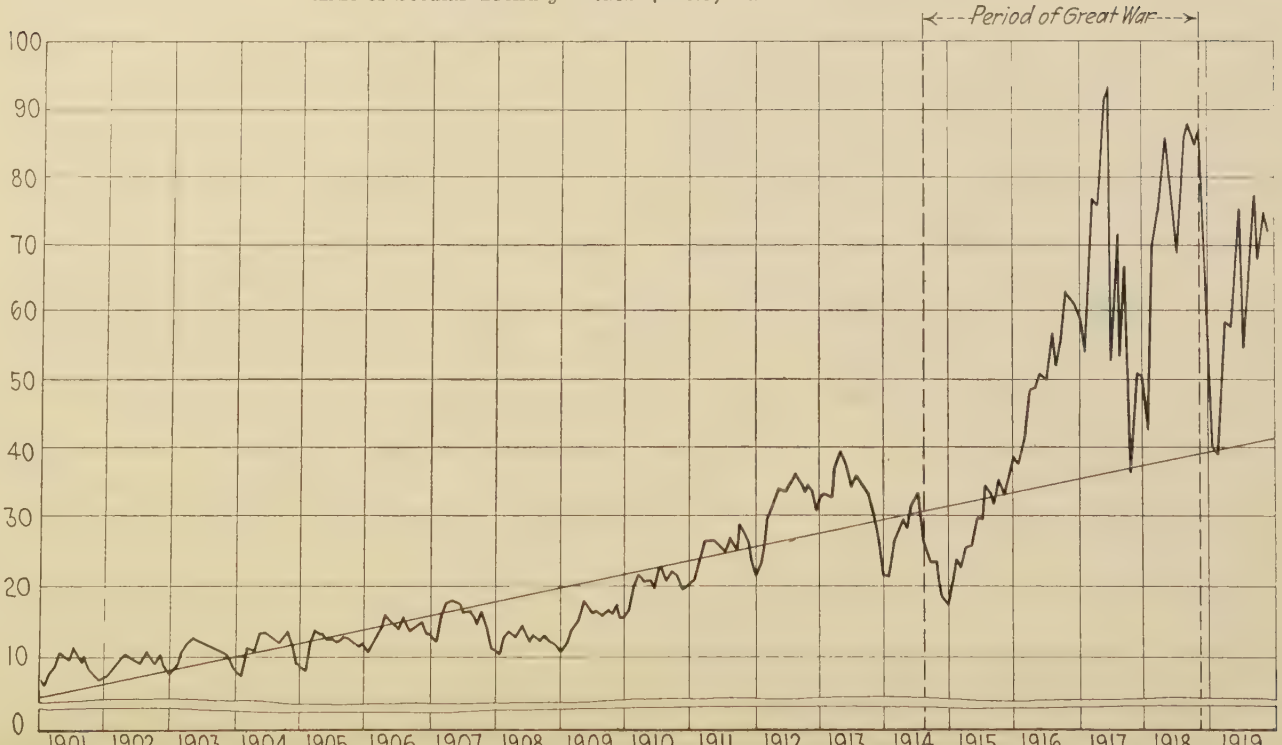


CHART 58.

This formula is derived from the regression equation as found on page 222.

$$b_2 = r \times \frac{\sigma_y}{\sigma_x},$$

in which b_2 = the slope of the line.
r = the coefficient of correlation.
σ_y = standard deviations of the actual items.
σ_x = standard deviations of time expressed in natural numbers.

This formula when expanded equals

$$b_2 = \frac{\Sigma xy}{n\sigma_x\sigma_y} \times \frac{\sigma_y}{\sigma_x} \qquad (1)$$

$$= \frac{\Sigma xy}{n(\sigma_x)^2}.$$

Since $\sigma^2 = \frac{\Sigma fd^2}{n}$

When f or frequency equals 1, then

$$\sigma^2 = \frac{\Sigma d^2}{N} \text{ or } \frac{\Sigma x^2}{N}$$

when x represents deviations of the years in natural numbers from the mean year.

By substitution of this value in Eq. (1)

$$b_2 = \frac{\Sigma xy}{n\frac{\Sigma_x{}^2}{n}} = \frac{\Sigma xy}{\Sigma x^2}. \qquad (2)$$

In this equation as applied to secular trend,

b = the slope of the line or normal increase.
x = the natural numbers assigned to the years as measured from the mean year.
y = the annual value of the items.

The division of Σxy by Σx^2 results in a value which measures the annual increase as well as the ratio of the respective sides as measured in terms of the tangent of the angle. A study does not, however, usually involve the entire period of growth of any industry, but selects some particular period for investigation, the beginning of which may have large production figures. The question then arises: At what year and from what figure should the growth element be added or subtracted? In other words: From what point is the tangent of the angle to be measured so that the line may have not only the right inclination but also

the proper elevation? If the figure is to be devoid of cyclic fluctuations any arbitrary year cannot be chosen, since the total for any single year may have a considerable cyclic element. Essentially a mean is required which will give weight to *both* prosperity and depression and so eliminate to a marked extent the cyclical variation. Since in correlation the tangent is measured from the intersection of the two means the logical point or year from which to add or subtract the growth element will in the problem be the *mean year* of the period used in determining the growth. This period, as already indicated, should be at least that of the cycle and preferably as many years as homogeneity of conditions permit. *The figure or value* for the mean year will likewise logically be the *mean of the sum of the yearly totals* for the number of years used in the computation, or by formula $\frac{\Sigma y}{N}$.

By obtaining the *mean sum* for the *mean year* of the years used in deriving the growth element a center is found for the data, from which may be added or subtracted the growth element, according as the deviation (x) from the year is plus or minus. This value depends upon the size of the deviation. The addition to or subtraction from the mean value now gives the secular trend (Y) for the given year.

Therefore, when x deviation is O, $Y = \frac{\Sigma y}{n}$.

The annual growth element was found by Eq. (2) to be $\frac{\Sigma xy}{\Sigma x^2}$. Therefore, Y for any year equals the mean value plus or minus the growth element multiplied by the number of years that the given year deviates from the mean year, or by formula

$$Y = \frac{\Sigma y}{n} + \frac{\Sigma xy}{\Sigma x^2} \times x$$

For simplification and in order to conform to the symbols customarily used let

Y = the secular trend.

$m = \frac{\Sigma xy}{\Sigma x^2}$ or normal growth.

$b = \frac{\Sigma y}{n}$ the mean of the sum of the annual totals.

x = deviation in years, expressed in natural numbers,

Therefore,

$$Y = mx + b.$$

By the substitution of various values for x the secular trend for any year may be derived.

TABLE LXXIV.—COMPUTATION OF SECULAR TREND OF ANNUAL EXPORTS FROM THE UNITED STATES TO CANADA, 1901–1913
Units of $1,000,000
Data taken from Table LXX

Year A	Value y B	x C	x^2 D	xy E		Secular trend F
1901	106	−6	36	− 636		66.6
1902	110	−5	25	− 550		89.4
1903	132	−4	16	− 528		112.2
1904	137	−3	9	− 411		135
1905	145	−2	4	− 290		157.8
1906	169	−1	1	− 169	−2,584	180.6
1907	187	0	0	.....		203.4
1908	153	1	1	153		226.2
1909	188	2	4	376		249
1910	242	3	9	726		271.8
1911	299	4	16	1,196		294.6
1912	376	5	25	1,880		317.4
1913	401	6	36	2,406	+6,737	340.2
Total	2,645	...	182	.....	4,153	

Method of Least Squares

Arithmetic mean of $y = \frac{2,645}{13} = 203.4 = b$

Slope or growth $\frac{\Sigma xy}{\Sigma x^2} = \frac{4,153}{182} = 22.8 = m$

By substitution in the equation $Y = mx + b$

$$Y = 22.8x + 203.4.$$

Table LXXIV is self-explanatory. The mean annual value $\frac{\Sigma y}{n}$ was the secular trend assigned to the year 1907, it being the mean year of the period used. The other secular trend values in column F were derived by either adding or subtracting, according to sign, the growth element, 22.8, for *every year* that any

given year exceeds or falls below the mean year, 1907, *i.e.*, the year 1913 exceeds the mean year by six years. Its secular trend, therefore, equals

$$Y = 22.8 \times 6 + 203.4 = 340.2.$$

The year 1901 falls below 1907 by six years, so that the secular trend for 1901 equals

$$Y = 22.8 \times -6 + 203.4 = 66.6.$$

The graphic presentation of the line of secular trend for yearly totals is found in Chart 57. All that is required in charting any line of secular trend, if it be linear, is the derivation of two values for Y. These values are then plotted on the ordinate for the year for which they were derived. A straight line drawn though these two points presents the lines of secular trend. The line on Chart 57 was derived as follows:

Given $Y = 22.8x + 203.4$

When $x = -5$ or year 1902

$$Y = 22.8 \times -5 + 203.4 = 89.4$$

when $x = +5$ or year 1912

$$Y = 22.8 \times 5 + 203.4 = 317.4$$

A + was placed on the ordinate for 1902 at the value 89.4, likewise on 1912 at 317.4. A straight line was then drawn through these points covering the entire period plotted. In any graphic presentation of the line of secular trend, derived by this method, the known values for the equation $Y = mx + b$ should be placed on the graph. By consulting Chart 57 the divergence between the method of the moving average and that of least squares becomes apparent.

Application to Monthly Data.—The method has been applied to yearly totals in which $Y = mx + b$. The data, however, may be recorded by months and the problem demand monthly analysis. Naturally the same values for m and b would not apply to the monthly figures. A new b must be derived for the value of *the mean month of the mean year*. The easiest method of deriving the monthly secular trend is to compute first the annual secular trend as given by the equation $Y = mx + b$. As a month is one-twelfth of a year, monthly values for m may be derived by dividing the m for annual values by $(12)^2$, the 12 being squared because the deviations are squared in the formula $\frac{\Sigma xy}{\Sigma x^2}$ which yields m. Therefore for monthly values the growth

element is $\frac{m}{144}$. Values for b are obtained by dividing the annual b by 12 $= \frac{b}{12}$.

The formula for monthly ordinates thus becomes

$$Y = \frac{mx}{144} + \frac{b}{12}$$

in which m and b are the yearly figures as already derived for the problem. But it should be remembered that x *is a monthly deviation from the mean of the months of the mean year.*

Application of Monthly Method to Problem.—Given $Y = 22.8x + 203.4$ for yearly data with 1907 as year of origin.

Therefore, for monthly figures

$$Y = \frac{22.8x}{144} + \frac{203.4}{12} = \underline{0.16x + 16.9}.$$

The equation indicates each month to deviate 0.16 from the preceding month, the deviations commencing from the mean of the twelve months of the year 1907.

Reference to Chart 58 will show that the data were plotted on lines, each line representing one month. Since there are twelve months in the year the mean falls midway between the sixth and seventh line. / / / / / /// / / / / /, *i.e., between June and July for the year* 1907. Therefore, to find the June ordinate for 1907, x, the deviation from the origin, is -0.5, while that for July is $+0.5$.

1 2 3 4 5 6 7 8 9 10 11 12

When $x = -0.5$ for June 1907

$$Y = 0.16 \times -0.5 + 16.9 \qquad = \underline{16.82}.$$

When $x = +0.5$ for July 1907

$$Y = 0.16 \times 0.5 + 16.9 \qquad = \underline{16.98}.$$

The values for all the other months of the study are given in Table LXXVI, column D, page 312.

The graphic presentation of the line of secular trend as shown in Chart 58 was derived in a manner similar to that applied to the yearly figures.

Given $Y = 0.16x + 16.9$.

By using June, 1907, as the basis of computation, in order to avoid fractional deviations,

$$Y = 0.16x + 16.82.$$

When $x = +60$ for June, 1912,

$$Y = 0.16 \times 60 + 16.82 = 26.42.$$

When $x = -60$ for June, 1902,

$$Y = 0.16 \times -60 + 11.82 = 7.22.$$

The line drawn through these points, as plotted on Chart 58, gives the line of secular trend.

150. Limitations of Method of Least Squares.—The method of least squares is the one usually used in the analyses of series which permit a straight-line fit. Some series are curvilinear. In such cases the compound-interest or parabola curve should be applied.

In order to obtain greater accuracy in the method of least squares, some writers have advocated a modification of the method usually used, advocating that the straight line be fitted with reference to *percentage or ratio deviations*. As one writer[1] states:

> The question arises whether it would not be preferable to fit the line so that the sum of the squares of these *ratios*, rather than the sum of the squares of the *actual* deviations, is a minimum.

The article concludes that such a method is preferable. This contention, however, has been disputed in that the method is more laborious and does not of necessity yield more accurate results.[2]

151. *Summary.*—There are various methods of computing secular trend, but the one which should be used is that of the line of least squares in which

$$Y = mx + b \text{ for yearly data}$$

and

$$Y = \frac{mx}{144} + \frac{b}{12} \text{ for monthly data.}$$

This method possesses greater ease and accuracy of computation than the moving average method and at the same time is similar to the method of moments. However, this method cannot be applied to series possessing curvilinear trends.

[1] Crum, W. L., "The Determination of Secular Trend," *Amer. Stat. Pub.*, p. 210, June, 1922.

[2] See article by Holbrook, W., *Amer. Stat. Pub.*, pp. 497–502, December, 1922.

References

CRUM, W. L.: "The Determination of Secular Trend," *Quart. Pub. Amer. Stat. Assoc.*, June, 1922.

DAVIES, G. R.: "Introduction to Economic Statistics," pp. 100–109.

FRANK, L. K.: "Long Time Trends," *Quart. Pub. Amer. Stat. Assoc.*, September, 1923.

JORDAN, D. F.: "Business Forecasting," p. 90.

KING, W. I.: "Elements of Statistical Method," pp. 191–192.

PRESCOTT, RAYMOND B.: "The Law of Growth in Forecasting Demand," *Quart. Pub. Amer. Stat. Assoc.*, December, 1922.

SECRIST, H.: "An Introduction to Statistical Methods."

YULE, G. U.: "An Introduction to the Theory of Statistics," chap. IX.

VANCE, RAY: "Business and Investment Forecasting," chap. III. *Rev. Econ. Sta.*, vol. I, pp. 8–18.

Problems

1. Why is there a tendency for the physical volume of the production of staple articles to increase over a period of time?

2. Do all series with the lapse of time indicate growth? Explain.

3. Under what conditions, if any, may secular trend be used as a forecaster?

4. Secular trend figures should be free from any cyclic or seasonal influence. In the computation of secular trend how may these two be eliminated?

5. How may secular trend be measured?

6. Plot the yearly and monthly data of business failures (problem 11, chap. XXIV).

7. For the yearly totals above determine the secular trend by

a. Moving average method.

b. Line of least squares.

8. Illustrate on graph for problem 6 the lines of secular trend as derived in problem 7.

9. Determine the monthly secular trend figures for the monthly data of problem 11, Chap. XXIV.

10. What are the limitations of the method of the line of least squares?

CHAPTER XXVI

CYCLIC FLUCTUATIONS

152. Introduction.—Cyclic fluctuations are those irregular but intermittent changes which take place in industry every few years. They are of unequal duration and occur with no definite periodicity and may be attributed to the cycle of business as exhibited by the periods of crisis, depression, recovery, and prosperity. The causes for these changes are still much-debated points. The phenomena, however, are of extreme importance to the business man.

Seasonal as well as secular changes command the attention of the business man, but he lives in fear of those periods of the cycle known as crisis and depression. The period of depression brings in its train the dislocation of business. The effect, however, of the phases of the business cycle is not the same for all industries and trades. Some businesses shows immediate response to changed business conditions, while others lag. Likewise the duration of the cycle varies for different industries. In other words, practically every business has a cyclic movement, but all do not synchronize with regard to amplitude, duration, and time of change. In constructing an index measuring cyclic changes in the production of the mines it was found that[1]

> Of the minerals comprising our final index, copper, zinc, and iron ore respond quickly and violently to changes in business prosperity; bituminous coal is fairly sensitive to general business conditions, while anthracite coal, petroleum, and silver are much less influenced. The metals, as a whole, exhibit a wide range of cyclical fluctuations than do the non-metals, which are dominated by the fuel group.

The divergences between industries indicate the necessity for devising some means of measuring cyclical changes, not merely with the idea of measuring past changes but rather to use the divergencies between different industries as a means of forecasting cyclical changes. By the construction of business

[1] Boody, E., "Cyclical Fluctuations in the volume of Mining, 1913–1923," *Rev. Econ. Stat.*, p. 77, April, 1924.

barometers and the derivation of the relationship of any particular industry to that barometer, the business man may have the means not of preventing the period of depression in his own business, but lessening the severity of it. The business man must, therefore, work out the cyclical variations in his own business and then by comparison with the general business cycle he may derive a forecaster of conditions. Some means therefore should be devised for the measuring of the cyclical element.

153. Normal Production.—In that two of the three commensurate fluctuations are already known, the methods having been developed in the previous chapters, the way is paved for the determination of the third. Since seasonal variation is that fluctuation devoid of both secular and cyclical variations and is expressed by indices, and secular trend is the growth element with seasonal and cyclical elements eliminated, the product of these two would give a *normal figure* of production, that is, what should normally be produced for any given month of any given year.

154. Moving Average Method.—Actual Figures.—The *normal figure* (seasonal index × secular trend) for any month, therefore, affords the base from which may be measured the differences in *actual* production. If the actual production exceeds or falls below what should normally be produced, this excess or deficit may be attributed to two causes:

1. The influence of the business cycle on production.
2. Changes due to abnormal conditions.

The latter are hardly commensurable, so that the excess is usually called cyclical. The student should, therefore, remember that the so-called cyclical changes may contain elements due to *abnormal conditions.* Abnormal conditions are usually known when they exist, so that misinterpretation is highly improbable. In graphic presentation it is well to indicate the abnormal periods (see Chart 60).

If annual totals are used instead of monthly data, the seasonal element does not enter into the problem, the normal is then represented by the secular trend, which when deducted from the original data results in the cyclical variations which are expressed in absolute numbers, the numbers being plus when the data exceed the normal and minus when they are below. These numbers when plotted above (+) and below (−) the *normal* or zero line the (years being placed below the abscissa) result in the

graphic presentation of the cyclical fluctuations. As an illustration of the derivation and graphic presentation of cyclic variations derived from annual figures by means of the moving average, Table LXXV and Chart 59 are given:

TABLE LXXV.—TABLE SHOWING CYCLIC FLUCTUATIONS IN COMMODITY *X*

Year	*X*	Secular trend by five-year average	Cyclic changes	Index of cyclic changes
1900	10			
1901	12			
1902	13	14	−1	93
1903	18	16	+2	112
1904	16	18	−2	89
1905	20	19	+1	105
1906	24	19	+5	126
1907	17	21	−4	81
1908	18			
1909	25			

If, however, monthly data are included in the study, one method of eliminating the seasonal element is to adopt the moving average method in the derivation of secular trend, the periods used being twelve months. This average constitutes the *normal*, from which are computed the cyclical changes.[1]

The moving average method, however, is not suitable for the derivation of secular trend. Therefore, the cyclic changes should be measured by using the method of least squares in finding the growth element from which may be computed the cyclical fluctuations.

155. Method of Least Squares.—If the data are annual figures all that is required is to subtract the secular trend (derived by least squares) from the original figures. *Care must be exercised in noting the signs, production below normal being minus, that above plus.*

Monthly data introduce the seasonal element. Therefore the normal for any one month requires that the secular trend be corrected for the seasonal change for that particular month.

[1] For illustration of method, see SECRIST, H., "An Introduction to Statistical Methods." (The moving average used in this study was thirteen months.)

The secular trend figures are absolute, while those of the seasonal variations are relative. The *normal* is obtained by multiplying the secular trend by the seasonal index for the required month and year, *i.e.*, the *normal* for January, 1924 = secular trend January, 1924 × seasonal variation January, 1924. This product when deducted from the original results in the cyclical variations being expressed in absolute numbers.

156. Application of Methods to the Problem.—Table LXXVI presents the method used in deriving the cyclical variations for

Chart of Cyclic Changes of Table LXXV

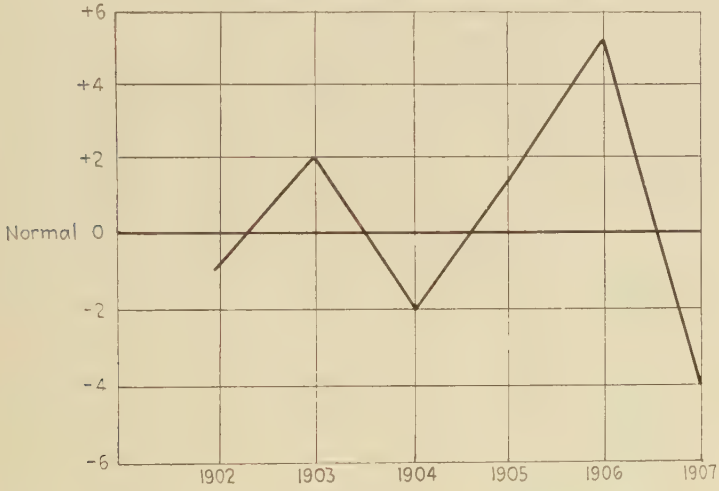


Chart 59.

exports from the United States to Canada. Column B of the table records the actual data as exhibited on Table LXX. Column C indicates the seasonal indices as derived in Chap. XXIV. Each January, irrespective of year, has a seasonal index of 84, while February has 82, etc. Column D was derived from Chap. XXV, in which for monthly data $Y = 0.16x + 16.9$ with the mean of the months of 1907 as the origin; June, 1907, as already indicated has a deviation of $-0.5x$, so that its secular trend equaled 16.682. The values recorded for the other months were derived by deducting 0.16 from the June, 1907, figure for every month that any month of any one year was below the June, 1907, origin. Likewise, adding 0.16 for those above June, 1907 (the figures were obtained by adding 0.16 to each preceding figure which had been corrected to the June base).

TABLE LXXVI.—TABLE SHOWING SEASONAL, SECULAR, AND CYCLICAL FLUCTUATIONS IN EXPORTS FROM THE UNITED STATES TO CANADA, 1901–1919

Units of $1,000,000

Year	Month	Data	Seasonal, per cent	Secular trend	Normal C × D	Difference B − E	F as per cent of E	G as in terms of its standard deviation
	A	B	C	D	E	F	G	H
1901	January	7.3	84	4.50	3.8	3.5	92.1	+3.61
	February	6.4	82	4.66	3.8	2.6	63.3	+2.48
	March	7.7	99	4.82	4.8	2.9	60.4	+2.37
	April	8.4	109	4.98	5.4	3.0	55.5	+2.17
	May	10.8	114	5.14	5.9	4.9	83.0	+3.25
	June	10.3	108	5.30	5.8	4.5	77.7	+3.04
	July	9.9	103	5.46	5.6	4.3	76.8	+3.04
	August	11.3	109	5.62	6.1	5.2	85.2	+3.36
	September	9.3	99	5.78	5.7	3.6	63.1	+2.47
	October	10.1	105	5.94	6.2	3.9	62.9	+2.46
	November	8.2	100	6.10	6.1	2.1	34.4	+1.35
	December	7.1	88	6.26	5.5	1.6	29.0	+1.17
1902	January	6.8	84	6.42	5.4	1.4	25.9	+1.01
	February	7.3	82	6.58	5.4	1.9	35.1	+1.37
	March	8.7	99	6.74	6.6	2.1	31.8	+1.24
	April	9.8	109	6.90	7.5	2.3	30.7	+1.20
	May	10.4	114	7.06	8.0	2.4	30	+1.19
	June	9.8	108	7.22	7.8	2.0	25.8	+1.01
	July	9.4	103	7.38	7.6	1.8	23.7	+0.93
	August	10.9	109	7.54	8.2	2.7	32.9	+1.29
	September	9.9	99	7.70	7.6	2.3	30.1	+1.19
	October	10.1	105	7.86	8.2	1.0	23.1	+0.90
	November	9.2	100	8.02	8	1.2	15	+0.58
	December	8.1	88	8.18	7.2	0.9	11.1	+0.43
1903	January	7.8	84	8.34	6.9	0.9	13.5	+0.52
	February	9.2	82	8.50	6.9	2.3	33.3	+1.31
	March	10.9	99	8.66	8.6	2.3	26.7	+1.04
	April	12.4	109	8.82	9.6	2.8	29.1	+1.14
	May	12.8	114	8.98	10.2	2.5	24.5	+0.96
	June	12.3	108	9.14	9.9	2.4	24.2	+0.96
	July	12.7	103	9.30	9.6	3.1	32.3	+1.26
	August	11.9	109	9.46	10.3	1.6	15.5	+0.60
	September	11.3	99	9.62	9.5	1.8	18.8	+0.74
	October	10.9	105	9.78	10.3	0.6	5.8	+0.23
	November	10.5	100	9.84	9.9	0.6	6	+0.23
	December	9.1	88	10.10	8.8	0.3	3.4	+0.3
1904	January	7.7	84	10.26	8.6	−0.9	−14.5	−0.56
	February	7.6	82	10.42	8.5	−0.9	−14.4	−0.56
	March	11.7	99	10.58	10.4	+1.3	12.5	+0.49
	April	11.1	109	10.74	11.7	−0.6	− 5.1	−0.20
	May	13.5	114	10.90	12.4	+1.1	8.8	+0.34
	June	13.6	108	11.06	11.9	1.7	14.3	+0.56
	July	12.8	103	11.22	11.5	1.3	11.2	+0.44
	August	12.3	109	11.38	12.3	0	0	0
	September	12.7	99	11.54	11.4	1.3	11.3	+0.44
	October	13.8	105	11.70	12.3	1.5	12	+0.46
	November	11.5	100	11.86	11.8	−0.3	− 3.4	−0.13
	December	9.0	88	12.02	10.6	−1.6	−15.1	−0.59
1905	January	9.1	84	12.18	10.2	−1.1	−10.7	−0.42
	February	8.0	82	12.34	10.1	−2.1	−20.7	−0.91
	March	12	99	12.50	12.4	−0.4	− 3.2	−0.13
	April	13.8	109	12.66	13.8			
	May	13.7	114	12.82	14.6	−0.9	− 6.1	−0.23
	June	12.7	108	12.98	14	−1.3	− 9.2	−0.36
	July	12.7	103	13.14	13.5	−0.8	− 5.9	−0.23
	August	12.6	109	13.30	14.5	−1.9	−13.3	−0.52
	September	12.7	99	13.46	13.3	−0.6	− 4.5	−0.18
	October	12.9	105	13.62	14.3	−1.4	− 9.4	−0.37
	November	12.7	100	13.78	13.8	−1.1	− 8.7	−0.34
	December	11.8	88	13.94	12.2	−0.4	− 3.2	−0.12

TABLE LXXVI.—TABLE SHOWING SEASONAL, SECULAR, AND CYCLICAL FLUCTUATIONS IN EXPORTS FROM THE UNITED STATES TO CANADA, 1901–1919 (*Continued*)

Units of $1,000,000

Year	Month	Data	Seasonal, per cent	Secular trend	Normal C × D	Difference B − E	F as per cent of E	G as in terms of its standard deviation
	A	B	C	D	E	F	G	H
1906	January	12.3	84	14.10	11.8	+0.5	4.1	+0.16
	February	10.8	82	14.26	11.7	−0.9	− 7.7	−0.30
	March	13.1	99	14.42	14.3	−1.2	− 8.3	−0.32
	April	14.4	109	14.58	15.9	−1.5	− 9.4	−0.37
	May	15.8	114	14.74	16.8	−1	− 5.9	−0.23
	June	14.9	108	14.90	16.1	−1.2	− 7.4	−0.29
	July	14.6	103	15.06	15.5	−0.9	− 5.7	−0.22
	August	15.6	109	15.22	16.6	−1	− 6	−0.23
	September	13.8	99	15.38	15.2	−1.4	− 9.2	−0.36
	October	14.7	105	15.54	16.3	−1.6	− 9.8	−0.39
	November	15	100	15.70	15.7	−0.7	− 4.4	−0.17
	December	13.6	88	15.86	13.9	−0.3	− 2.1	−0.08
1907	January	13.6	84	16.02	13.5	+0.1	+ 0.7	+0.03
	February	12.4	82	16.18	13.3	−0.9	− 7	−0.27
	March	15.9	99	16.34	16.1	−0.2	− 1.2	−0.05
	April	17.8	109	16.50	18.0	−0.2	− 1.2	−0.05
	May	18	114	16.66	18.9	−0.9	− 5.4	−0.21
	June	17.9	108	16.82	18.1	−0.2	− 1.2	−0.05
	July	16.5	103	16.98	17.5	−1	− 5.9	−0.23
	August	16.8	109	17.04	18.6	−1.8	−10.9	−0.43
	September	15	99	17.20	17	−1.5	− 8.7	−0.34
	October	16.8	105	17.36	18.2	−1.4	− 6.2	−0.24
	November	14.6	100	17.52	17.5	−2.9	−16.6	−0.65
	December	11.2	88	17.68	15.5	−4.3	−27.7	−1.08
1908	January	11.1	84	17.84	15	−3.9	−26	−1.02
	February	10.7	82	18	14.8	−4.1	−27.9	−1.09
	March	13.4	99	18.16	18	−4.6	−25.5	−1
	April	14.0	109	18.32	19.9	−5.9	−29.7	−1.16
	May	13.1	114	18.48	21	−7.9	−37.6	−1.47
	June	13.8	108	18.64	20.1	−6.3	−31.3	−1.23
	July	12.9	103	18.80	19.4	−6.5	−33.5	−1.31
	August	13	109	18.96	20.6	−7.6	−36.4	−1.43
	September	12.6	99	19.12	18.9	−6.3	−33.3	−1.30
	October	13.3	105	19.28	20.2	−6.9	−34.1	−1.34
	November	12.4	100	19.44	19.4	−7	−34	−1.33
	December	12.1	88	19.60	17.2	−5.1	−29.6	−1.16
1909	January	10.8	84	19.76	16.6	−5.8	−34.9	−1.37
	February	12.2	82	19.92	16.3	−4.1	−25.1	−0.98
	March	14.4	99	20.08	19.9	−5.5	−27.7	−1.09
	April	15.2	109	20.24	22	−6.8	−30.9	−1.21
	May	17.7	114	20.40	23.2	−5.5	−23.7	−0.93
	June	16.8	108	20.56	22.2	−5.4	−24.3	−0.95
	July	16.8	103	20.72	21.3	−4.5	−21.1	−0.83
	August	16.7	109	20.88	22.8	−5.1	−22.4	−0.88
	September	17.0	99	21.04	20.8	−3.8	−18.2	−0.71
	October	16.9	105	21.20	22.2	−5.3	−23.8	−0.93
	November	17.7	100	21.36	21.3	−3.6	−16.8	−0.65
	December	15.7	88	21.52	18.9	−3.2	−16.9	−0.65
1910	January	15.4	84	21.68	18.2	−2.8	−15.4	−0.60
	February	16.3	82	21.84	17.8	−1.5	− 8.4	−0.33
	March	20.2	99	22	21.8	−1.6	− 7.3	−0.28
	April	21.9	109	22.16	24	−2.1	− 8.7	−0.33
	May	20.8	114	22.32	25.4	−4.6	−18.1	−0.71
	June	20.7	108	22.48	24.3	−3.6	−14.8	−0.58
	July	19.9	103	22.64	23.3	−3.4	−14.6	−0.57
	August	22.9	109	22.80	24.8	−1.9	− 7.6	−0.30
	September	26.7	99	22.96	22.7	−2	− 8.8	−0.34
	October	22	105	23.12	24.2	−2.2	− 9	−0.35
	November	21.6	100	23.28	23.2	−1.6	− 6.4	−0.25
	December	19.8	88	23.44	20.6	0.8	− 3.4	−0.13

TABLE LXXVI.—TABLE SHOWING SEASONAL, SECULAR, AND CYCLICAL FLUCTUATIONS IN EXPORTS FROM THE UNITED STATES TO CANADA, 1901–1919 (*Continued*)

Units of $1,000,000

Year	Month	Data	Seasonal, per cent	Secular trend	Normal C × D	Difference B − E	F as per cent of E	G as in terms of its standard deviation
	A	B	C	D	E	F	G	H
1911	January	20.1	84	23.60	19.8	+ 0.3	+ 1.5	+0.06
	February	20.8	82	23.76	19.5	+ 1.3	+ 6.7	+0.20
	March	23.5	99	23.92	23.6	0.1	0.4	+0.01
	April	26.3	109	24.08	26.3	0	0	0
	May	26.3	114	24.24	27.6	− 1.3	− 0.5	−0.02
	June	25.9	108	24.40	26.3	+ 0.4	+ 1.5	+0.06
	July	24.8	103	24.56	25.3	+ 0.5	− 0.9	−0.07
	August	27.1	109	24.72	26.9	+ 0.2	+ 0.9	+0.03
	September	25.8	99	24.88	24.6	0.8	3.2	+0.12
	October	28.9	105	25.04	26.3	2.6	9.8	+0.35
	November	26.6	100	25.20	25.2	1.4	5.5	+0.21
	December	23.1	88	25.36	22.3	0.8	3.6	+0.14
1912	January	21.3	84	25.52	21.4	− 0.1	− 0.4	−0.01
	February	23.3	82	25.68	21	+ 2.3	+10.9	+0.40
	March	29.6	99	25.84	25.5	+ 4.1	+16	+0.62
	April	31.0	109	26.0	28.4	+ 2.6	+ 9.1	+0.35
	May	34.0	114	26.16	29.7	+ 3.3	+11.1	+0.43
	June	33.7	108	26.32	28.2	+ 5.5	+19.5	+0.76
	July	35.0	103	26.48	27.3	7.6	28.3	+1.11
	August	35.7	109	26.64	29.0	6.7	23.1	+0.90
	Sepremper	33.7	99	26.80	26.5	7.2	27.1	+1.06
	October	34.4	105	26.96	28.3	6.1	21.5	+0.83
	November	34.5	100	27.12	26.1	6.9	25.5	+1.00
	December	30.6	88	27.28	24	6.6	27.5	+1.08
1913	January	32.7	84	27.44	23	9.7	42.1	+1.65
	February	33.0	82	27.60	22.6	10.4	46	+1.80
	March	32.9	99	27.76	27.5	5.4	19.6	+0.77
	April	37.4	109	27.92	30.3	7.1	23.4	+0.92
	May	39.4	114	28.08	31.9	8.5	26.6	+1.04
	June	37.6	108	28.24	30.5	7.1	23	+0.90
	July	34.2	103	28.40	29.2	5	17	+0.66
	August	35.4	109	28.56	31.1	4.3	13.8	+0.54
	September	30.7	99	28.72	28.4	2.3	8.1	+0.32
	October	33.1	105	28.88	30.3	2.8	9.2	+0.36
	November	30.6	100	29.04	29	1.6	5.5	+0.21
	December	26	88	29.20	25.7	0.3	1.1	+0.05
1914	January	21.4	84	29.36	24.6	− 3.2	−13	−0.51
	February	21.4	82	29.52	24.2	− 2.8	−11.6	−0.45
	March	26.2	99	29.68	29.4	− 3.2	−10.9	−0.43
	April	28.2	109	29.84	32.5	− 4.3	−13.2	−0.51
	May	29.6	114	30.00	34.2	− 4.6	−13.4	−0.52
	June	28.2	108	30.16	32.6	− 4.4	13.5	−0.52
	July	31.6	103	30.32	31.2	+ 0.4	+ 1.2	+0.04
	August	33.5	109	30.48	33.2	0.3	0.9	+0.03
	September	26.0	99	30.64	30.3	− 4.3	−14.1	−0.55
	October	23.6	105	30.80	32.3	− 8.7	−26.9	−1.05
	November	23.5	100	30.96	30.9	− 7.4	−22.9	−0.90
	December	18.4	88	31.12	27.4	− 9	−32.4	−1.27
1915	January	17.4	84	31.28	26.3	− 8.9	−33.8	−1.32
	February	24.2	82	31.44	25.8	− 1.6	− 6.2	−0.24
	March	22.6	99	31.60	31.3	− 8.7	−27.7	−1.09
	April	25.6	109	31.76	34.6	− 9	−26	−1.01
	May	25.5	114	31.92	36.4	−10.9	30	−1.13
	June	29.6	108	32.08	34.6	− 5	−14.4	−0.56
	July	29.5	103	32.24	33.3	3.8	−11.4	−0.45
	August	34.8	109	32.40	35.3	− 0.5	− 1.4	−0.04
	September	31.9	99	32.56	32.2	− 0.3	− 0.9	−0.03
	October	35.3	105	32.72	34.4	+ 0.9	+ 2.6	+0.10
	November	32.6	100	32.88	32.8	− 0.2	− 0.6	−0.02
	December	36.0	88	33.04	29	+ 7	+24.1	+0.94

TABLE LXXVI.—TABLE SHOWING SEASONAL, SECULAR, AND CYCLICAL FLUCTUATIONS IN EXPORTS FROM THE UNITED STATES TO CANADA, 1901–1919 (*Continued*)

Units of $1,000,000

Year	Month	Data	Seasonal, per cent	Secular trend	Normal C × D	Difference B − E	F as per cent of E	G as in terms of its standard deviation
	A	B	C	D	E	F	G	H
1916	January	38.3	84	33.20	27.9	+10.4	+37.5	+1.47
	February	37.7	82	33.36	27.4	+10.3	+37.6	+1.47
	March	41.2	99	33.52	33.1	+11.1	+33.6	+1.32
	April	49	109	33.68	36.7	+12.3	+33.5	+1.31
	May	49.6	114	33.84	38.6	+11	+28.5	+1.11
	June	51.1	108	34.0	36.5	+14.6	+40	+1.57
	July	50.1	103	34.16	35.2	+14.9	+42.3	+1.65
	August	56.9	109	34.32	37.4	+19.5	+52.1	+2.04
	September	52.1	99	34.48	31.2	+20.9	+66.9	+2.62
	October	55.4	105	34.64	36.5	+18.9	+51.8	+2.02
	November	62.9	100	34.80	34.8	+28.1	+80.7	+3.16
	December	61.0	88	34.96	30.7	+30.3	+98.9	+3.88
1917	January	58.9	84	35.12	29.5	29.4	99	+3.88
	February	53.8	82	35.28	28.9	24.9	86.1	+3.37
	March	77.0	99	35.44	35.0	42	120	+4.70
	April	75.7	109	35.60	38.8	36.9	95.1	+3.73
	May	90.9	114	35.76	40.8	50.1	122.5	+4.80
	June	94.5	108	35.92	38.8	55.7	143.6	+5.63
	July	53.0	103	36.08	37.1	26	70.1	+2.71
	August	71.4	109	36.24	39.5	31.9	80	+3.13
	September	53.9	99	36.40	36.0	17.9	49.7	+1.95
	October	64.0	105	36.56	38.4	25.6	66.6	+2.61
	November	36.2	100	36.72	36.7	0	0	
	December	50.8	88	36.88	32.5	18.3	56.6	+2.22
1918	January	50.3	84	37.04	31.1	19.2	61.7	+2.41
	February	42.6	82	37.20	30.5	11.1	31.9	+1.25
	March	69.5	99	37.36	37	32.5	86.5	+3.39
	April	75.2	109	37.50	40.9	34.3	83.8	+3.28
	May	85.6	114	37.68	42.9	42.7	99.3	+3.89
	June	77.0	108	37.84	40.9	36.1	88.2	+3.46
	July	69.0	103	38.0	39.1	29.9	76.5	+3.00
	August	86.1	109	38.16	41.6	44.5	108.5	+4.25
	September	89.3	99	38.32	37.9	51.4	135.6	+5.31
	October	82.5	105	38.48	40.4	42.1	104.2	+4.09
	November	87.0	100	38.64	38.6	41.4	107.2	+4.20
	December	72.7	88	38.80	34.1	38.6	113	+4.43
1919	January	55.7	84	38.96	32.7	23.0	70.3	+2.75
	February	40.6	82	39.12	32.1	8.5	25.5	+1.05
	March	38.6	99	39.28	38.9	0.3	0.8	+0.03
	April	59.4	109	39.44	43.0	16.4	38.1	+1.49
	May	57.3	114	39.60	45.0	12.3	27.3	+1.07
	June	75.2	108	39.76	42.9	32.3	75.3	+2.95
	July	55.3	103	39.92	40.1	15.2	37.9	+1.48
	August	62.4	109	40.08	43.7	18.7	42.8	+1.68
	September	77.9	99	40.24	39.8	38.1	95.6	+3.74
	October	68.0	105	40.40	42.4	25.6	59.9	+2.35
	November	74.7	100	40.56	40.6	34.1	84	+3.29
	December	71.9	88	40.72	35.9	36	100	+3.92

Column E is the *normal* export figure, the product of the seasonal and secular fluctuations as in columns C and D. The figures in column E, when *subtracted from* the original data (column B), yield the changes due to cyclic and abnormal conditions. For general purposes, however, these are called cyclical variations. These changes are in absolute numbers and could be plotted in the same way as illustrated in Chart 59.

CHART OF CYCLIC CHANGES OF TABLE LXXV

CYCLIC FLUCTUATION OF EXPORTS FROM THE UNITED STATES TO CANADA, 1901–1919

Data in Table LXXVI

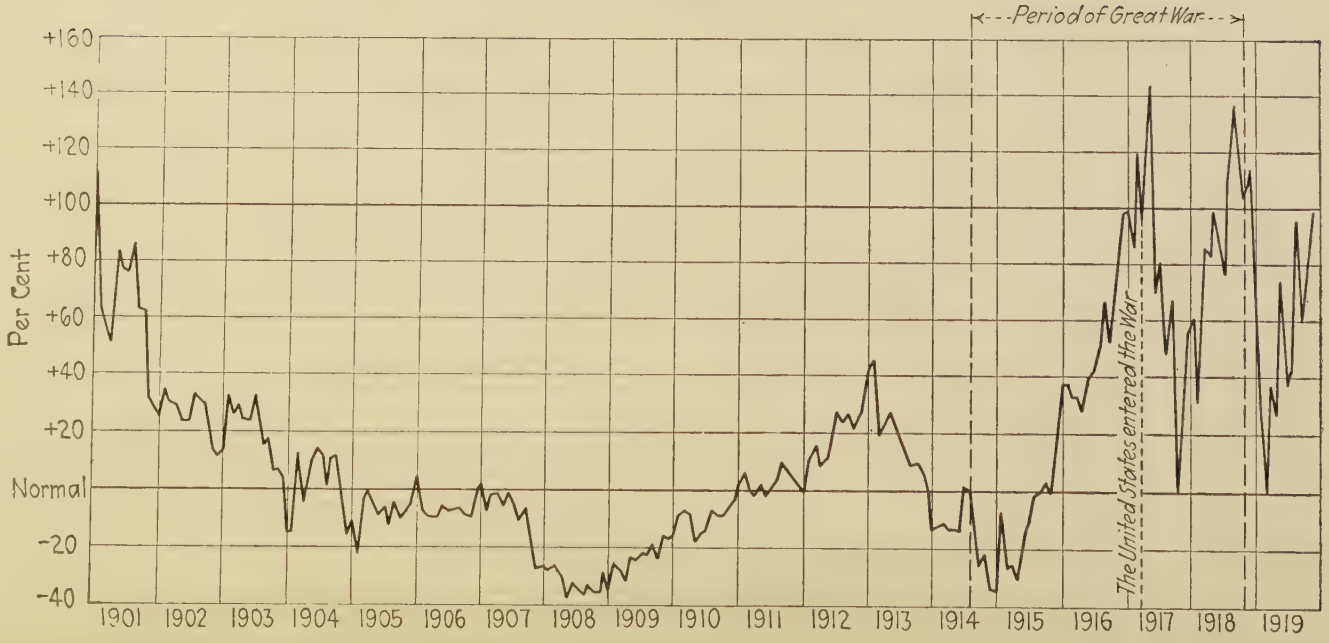
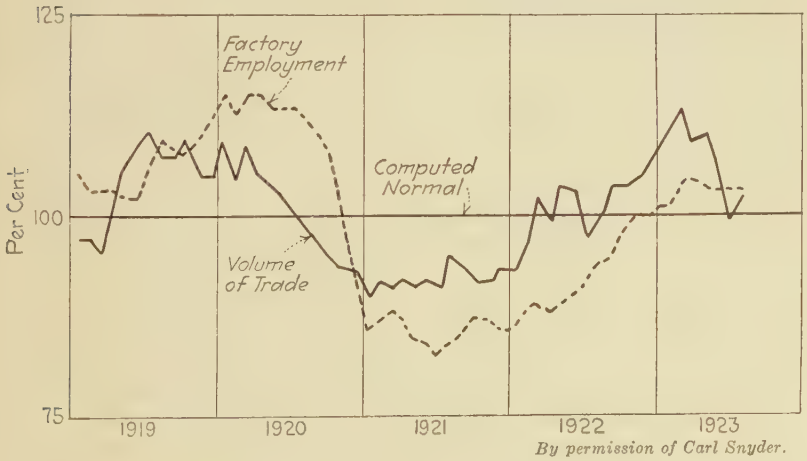


CHART 60.

157. Comparability of Cyclic Changes.—As already stated, the business man wishes to know not only the cycle of his own business but how that cycle relates itself to those of other trades. The figures of the variations, in themselves, may convey little, whereas if the two cycles to be compared were plotted on the same graph relationships may be exhibited. The cyclical variations, however, if expressed in absolute numbers may present difficulties in bringing the two curves into juxtaposition, in that the one series may be expressed in units of $1,000,000, while the other but $1,000 or 1,000 tons.

Homogeneity and a certain degree of coincidence of the two curves may be obtained by expressing the difference of the actual from normal as a percentage *of normal*. Normal is the production which would be found if there were a uniform growth corrected for the seasonal activity. If there were no seasonal variation, the secular trend figures would be the *normal* as well as the normal growth figures. Comparability may be obtained by dividing the difference from normal by the normal production.

In Table LXXVI, column G was obtained by dividing the difference as in column F by the normal, column E. The cyclical variations as expressed in column G are shown in Chart 60. The ease with which diverse series may be brought into juxtaposition by expressing all differences from normal as relatives rather than absolute numbers is illustrated in the chart below.



By permission of Carl Snyder.

[[Chart 61.

158. The Group Index.—The need for obtaining percentage deviations or relatives from normal, in contrast to recording cyclical fluctuations in absolute numbers, is emphasized by the fact that it is sometimes desirous to group related industries in order to obtain a group index. The object of such an index is to measure the changes from normal of the entire group rather than one particular industry. If absolute variations were given, the heterogeneous units of measurements would make the construction of such an index difficult. On the other hand, if the index is to measure amplitude of changes, the method of indices makes possible a greater degree of comparability.

The relatives for the individual commodities or industries entering into the group are usually weighted according to some predetermined purpose, as discussed in the section on Weighting. Each commodity relative, the percentage deviation from normal, for any given time or period of time, may be multiplied by its respective weight. These products when summated and divided by the total of weights used result in the group index. This index measures changes from normal, since the indices were expressed as percentages of normal. The geometric mean, however, is the average usually used in deriving the group index.

159. Summary.—Cyclic fluctuations as used in this chapter includes all those fluctuations from the *normal figure.* They, therefore, may include variations arising through abnormal conditions. These conditions, however, are usually known and should be indicated in the graphic presentation of the data (see Chart 60).

The *normal* figure from which these variations are measured is derived by multiplying the seasonal index by the secular trend for each and every month. The excess or deficit of the original data from this figure constitutes a plus or minus fluctuation.

If comparison of cycles or the construction of a group index is desired, the cyclic deviation may be expressed as relatives to their respective normal figures, the variations being divided by the respective normal.

Cyclic variations are important and form the basis of forecasting changes in basic business conditions.

The method of their derivation may be summarized briefly:

1. Obtain seasonal indices for respective months—link relative method.

2. Obtain secular trend—method of least squares.

3. Multiply seasonal index by related secular trend.

4. Subtract the product of (3) from the original data. The result is cyclic variations in absolute numbers.

5. Express the fluctuations, as in (4), as percentages of the normal, as in (3).

References

DAVIES, G. R.: "Introduction to Economic Statistics," pp. 123–126.
JORDAN, D. F.: "Business Forecasting," chap. II and III.
MITCHELL, WESLEY C.: "Business Cycles."
MOORE, H. L.: "Economic Cycles; Their Law and Cause."
SECRIST, H.: "An Introduction to Statistical Methods."
VANCE, RAY: "Business and Investment Forecasting," chap. III.
Rev. Econ. Stat., pp. 31–36, vol. I, 1919.

Problems

1. How would you derive a normal production figure for any series?

2. If actual data deviate from the normal, what are these deviations to be attributed to?

3. How does the business cycle affect business in general?

4. Should a business man accept the general business cycle as criterion of the fluctuations in his own business? Explain.

5. *a.* If a problem treats of yearly totals only, how could you determine the cyclic changes by the use of the moving average?

b. Apply the method to the yearly data for problem 11, chap. XXIV.

6. Given in Chap. XXIV, problem 11, the seasonal indices, and in Chap. XXV, problem 9 the secular trend for monthly business failures, compute the cyclical fluctuations for the series.

7. Represent graphically the results of problem 6.

8. *a.* If you wished to compare graphically the cyclical fluctuations of two series which were measured in units vastly different in magnitude, how could you bring the two curves into juxtaposition?

b. Illustrate the method in (*a*) by applying it to the results of problem 6 above.

CHAPTER XXVII

BUSINESS BAROMETERS

160. Introduction.—The object of this chapter is not to derive a business barometer nor to analyze all the points involved in its construction. For such information the student is referred to Brookmire's, Babson's, Moody's, and the *Review of Economic Statistics.* All of these services have constructed business barometers by methods which vary to a considerable degree. In this chapter will be discussed the need for and the fundamental principles involved in the construction of any business barometer.

161. The Effects of the Cycle.—The phenomena of the business cycle have already been referred to and the periods of depression leave memories of unemployment for the workingman, low prices for the retailers, scarcity of orders for the producer, and accumulation of funds for the banker. The phenomena of the phases of depression and recovery are summarized by the Harvard Committee on Economic Research as follows:

> Starting with phase 1, *depression*, we have, as a legacy from the preceding period of liquidation, low prices of securities, and small volume of speculation, depressed commodity prices and sluggish business activity, declining rates on commercial paper and increasing bank reserves. With marked dullness in speculation and business, the demand for credit slackens to a point which results in an accumulation of large quantities of money in the banks of the country. These funds seek an outlet in the security markets through the purchase of stocks and bonds, which are now fluctuating around their lowest levels in the cycle. Idle money, therefore, is drawn into the speculative market to purchase high-grade bonds, high-grade stocks, and the lower-grade stocks in the order named, and an upward movement in security prices and speculative activity is under way.
>
> After this upward swing in speculation has been in progress for several months business lifts its head from the depths of depression, commodity prices begin to advance, and we enter the second phase of the cycle, *business recovery.* This phase is characterized by advancing security prices, usually a strong bull market in stocks, rising commodity prices,

and increasing business activity. Presently, however, the demands of speculation for funds to finance heavy stock market operations while prices of stock are rising and the increasing demands of business for funds to facilitate industrial expansion cannot both be granted to the degree desired, and money rates begin to harden under the pressure of the demand for credit. At this point we pass into phase 3 of the cycle, *business prosperity*.[1]

162. The Need and Uses of a Business Barometer.—The dislocation of business in the depression period was quite apparent in the kaleidoscopic fall in prices during the latter part of 1920. Unemployment became so accute that President Harding called a Conference on Unemployment, which resulted in the publication of "Business Cycles and Unemployment" by the National Bureau of Economic Research, which analyzed the cycle and emphasized the need for forecasting changes in business conditions as it affects labor. A barometer is needed to record not only the humidity and temperature but also the probable weather conditions so that business may have its stock of umbrellas and rubbers on hand when the downpour comes.

Some of the uses to which a business barometer may be put are summarized as follows:

> Economists and sociologists need such a barometer when dealing with the phenomena of a dynamic society; government officials when handling the problem of unemployment or when considering the advisability of inaugurating large government undertakings; manufacturers and dealers when considering the desirability of making extensions to their plants or of contracting or expanding their purchases, sales, or committments; bankers need a business barometer to guide them in extending or calling their loans and discounts; investors need one to direct their purchases and sales of securities.[2]

163. The Construction of a Barometer.—*a. Its Basis.*—Reference has already been made to the effect of the business cycle on different industries; there are variations in time, duration, and amplitude in the occurrence of the various phases of the business cycle. The difference in time and duration forms the basis of the construction of a business barometer. As seen by the quotation on page 320, some forms of business enter

[1] "The Harvard Index of General Business Conditions," Its Interpretation (pamphlet, 1923), p. 4.

[2] PERSONS, WARREN M., "Construction of a Business Barometer Based upon Annual Data," *Amer. Econ. Rev.*, p. 739, December, 1916.

the phases of the business cycle in advance of others. If the members of the advance guard were assembled into homogeneous companies, the concerted action of all might permit the establishing of a line of communication with the main guard, so that the changing conditions of the enemy's front might be communicated back and preparations made accordingly. A group index of certain forms of business which move in advance of the cycle as it affects business in general may act as a barometer or forecaster, communicating by its rise or fall the change in business conditions.

Time and duration were the elements stated to be required in the construction of a barometer. Amplitude is omitted, to a marked degree, since the function of the forecaster is to indicate when depression, etc., is about to take place. Such indication, at once, affects, or rather should affect, amplitude, because business prepares for the change and so smooths out, to a certain extent, the peaks of prosperity and depression.

Those forms of business which exhibit the same time and practically the same duration of the occurrence of any one phase of the cycle should be grouped together. By the analysis of these various groups some may be found to respond more quickly to changing business conditions, and therefore may act as a forecaster for the less sensitive groups. The term "group" is used because no single industry or commodity can serve *adequately* as a barometer, because abnormal conditions, such as strikes, etc., in one industry, would affect the index, whereas by using a group these abnormal conditions may, in all probability, not prevail throughout the group. The problems therefore, in the construction of a barometer become:

1. The grouping of series according to the time of change.
2. Measuring the divergence of the time change between groups in order to derive a forecaster.

b. The Grouping of Series Having Similarity in Time Changes.— A barometer measures not normal growth but cyclical changes. Therefore, the association between series is that of cyclical fluctuation. The first problem becomes one of measuring the cyclical changes of the various series that are considered to be related to the cycle; then the comparison of these changes to determine if any series synchronize in their time of change.

(1) Comparability of Series by Means of Standard Deviation. The methods of measuring cyclical fluctuations were dis-

cussed in the preceding chapter and the fluctuations were expressed both as absolute deviations and percentage changes from *normal*. In a barometer, however, the interest centers not on amplitude but on the time of change. Since two series may have totally different percentage deviations from the normal, owing to differences in the units of measurement and the degree of dispersion, they are not strictly comparable. Comparability should be obtained by expressing the respective percentage deviations from normal in terms of the standard deviation of those percentage figures for any one series. When so expressed all series are reduced to the same units and, since six times the standard deviation generally covers 99 per cent of the frequency of any series, the series are reduced to graphic comparability, the curves being confined within small limits.

In order to illustrate this method it was applied to the problem. The percentage deviations as found in Table LXXVI column (G) were tabulated into class intervals in order to measure the degree of dispersion. The method is illustrated in Table LXXVII.

It should be observed that in Table LXXVII the dispersion is computed from the normal years 1901 to 1913. The war years are omitted because the inclusion of such years would give an abnormal degree of dispersion. In order to condense the percentage deviations a class interval of 5 per cent was determined upon and since there were both positive and negative deviations the interval was measured on both side of the zero. The data, as contained in column G, Table LXXVI, were then tabulated into the intervals as shown in Table LXXVII. Thirteen years were tabulated, which yielded a total frequency of 156, for which the dispersion was found to be 25.5 per cent.

By dividing the percentage deviations, column G, Table LXXVI by the standard deviation the differences were expressed in terms of the standard deviation, column H. The results of this method are shown in Table LXXVI, column H, January, 1901, percentage deviation of 92.1 when divided by the standard deviation of 25.5 per cent equals +3.61. (In this method care must be exercised in maintaining the signs).

(2) Grouping of Series Which Synchronize in Time Element. The cyclical variations as expressed in column H, Table LXXVI are shown graphically in Chart 62. A recollection of the properties of the standard deviation should make it evident that other series when so expressed may readily be superimposed

TABLE LXXVII.—COMPUTATION OF THE STANDARD DEVIATION OF THE PERCENTAGE DEVIATIONS FROM THE NORMAL OF EXPORTS FROM THE UNITED STATES TO CANADA, 1901–1913

Per cent deviation	Tabulation	Total frequency	d	df	d^2f
−35–40	//	2	− 8	−16	128
−30–35	𝍸 //	7	− 7	−49	343
−25–30	𝍸 ///	8	− 6	−48	288
−20–25	𝍸 /	6	− 5	−30	150
−15–20	𝍸 /	6	− 4	−24	96
−10–15	𝍸 ////	9	− 3	−27	81
− 5–10	𝍸 𝍸 𝍸 𝍸	25	− 2	−50	100
− 1– 5	𝍸 𝍸 ////	14	− 1	−14 −258	14
+ 0– 5	𝍸 𝍸 ///	13	0	0	0
+ 5–10	𝍸 𝍸	10	1	10	10
10–15	𝍸 𝍸	10	2	20	40
15–20	𝍸 //	6	3	18	54
20–25	𝍸 ////	9	4	36	144
25–30	𝍸 𝍸	10	5	50	250
30–35	𝍸 ///	8	6	48	288
35–40	/	1	7	7	49
40–45	//	2	8	16	128
45–50		0	9		
50–55		0	10		
55–60	/	1	11	11	121
60–65	////	4	12	48	576
65–70	—	0	13		
70–75		0	14		
75–80	//	2	15	30	450
80–85	/	1	16	16	256
85–90	/	1	17	17	289
90–95	/	1	18	18 +345	324
Total.....	.....	156	..	+ 87	4,179

$$\text{Standard deviation} = \sqrt{\frac{\Sigma fd^2x}{f} - K^2} \times C$$

$$\text{by substitution} = \sqrt{\frac{4,179}{156} - \left(\frac{345 - 258}{156}\right)^2} \times 5$$

$$= 25.5 \text{ standard deviation.}$$

upon the same graph, since they will cover approximately the same field. Comparability, therefore, is attained through the standard deviation.

In the construction of a barometer each of the selected series should be treated in a manner similar to that above. The degree of association between the series may then be *indicated* by graphic presentation. The cyclic variations in terms of the standard deviation of the various series, if plotted upon translucent paper, may be superimposed one upon another. Hence, by

CYCLIC FLUCTUATIONS OF EXPORTS FROM THE UNITED STATES TO CANADA, YEARS 1901–1919
Fluctuations in terms of the Standard Deviation
Data in Table LXXVI

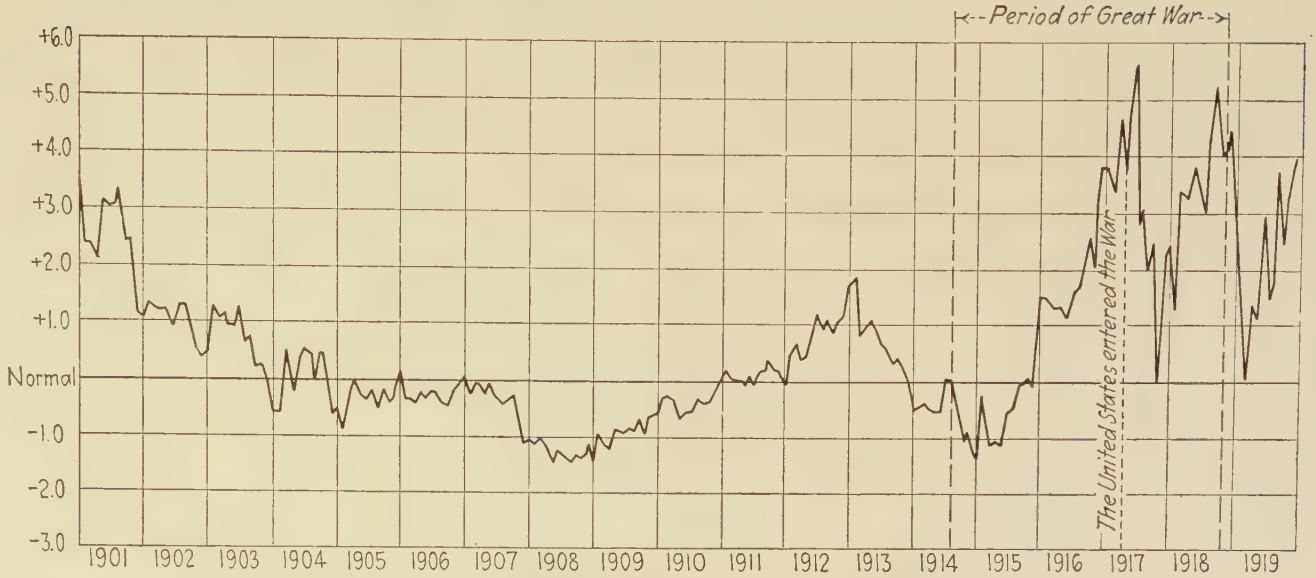


CHART 62.

inspection, the curves for the various data may be sorted into groups which synchronize in the time of movement or change. Those which indicate an upswing in January of any given year may constitute one group of series, while those which change in March may form another group.

Graphic methods, however, as already emphasized elsewhere, are only approximations. Therefore, the rough groupings should be submitted to mathematical analysis in order to confirm the assumed association between the series for any one group. The mathematical device for measuring the degree of association is correlation.

(3) Correlation of Time Series.—Correlation involves the sum of the products of the supposedly associated paired items divided by the product of the number of paired items and the respective standard deviations. By formula:

$$r = \frac{\Sigma xy}{n\sigma_x\sigma_y}.$$

When cyclical fluctuations are expressed in terms of the respective standard deviations the only unknown quantity in the correlation formula is Σxy. The number of paired items (n) is known, it being the number of months or years used in computation. The standard deviations of the respective series are known, since the cyclic fluctuations were derived from them. The unknown Σxy is obtained by summating the product of the paired items. The series within any one group were by graphic presentation supposed to synchronize in time of movement. Therefore, Σxy should be at its maximum when the same months of like years are multiplied together, *i.e.*, January, 1901 of X series by January, 1901, of Y series when both series belong to group Z.

(4) Lag.—The association if it be significant must be at least 0.5 and at least six times its probable error. Graphic determination makes it very difficult to tell if January of one series synchronizes with January or February of others within the group. The series therefore should be tested for lag, that is, to see if one series actually changes some period of time in advance of others, the one rising in January while others take the upswing at some later date. The test for lag consists of changing the association of paired items. Coefficients should be worked out for different associations, January of X series with February of Y series, or

with March of Y series, etc., or by reversing the relationship, January of Y series with February of X series (all the series belonging to the same group Z). Naturally, a different total Σxy will be obtained for the various pairing of items. Likewise (n) the number of paired items will also change. Therefore the value of r will not be constant. The association which yields the highest coefficient of correlation will indicate the probable relationship.

By applying this test to all series included in the respective rough groupings, they may be sorted in such a way that only those which synchronize in time of change, as evidenced by the coefficient of correlation being at its maximum when like months are correlated, are included in the final groupings.

The correlation test is not infallible, since it is only a mathematical device, but it is necessary in order to sort out the supposedly related series. By this process the raw material is sorted into groups of related series so that in each group there are series whose fluctuations are not identical in amplitude but only in *time of change*. The selected groups, naturally, have different periods of change and it is these divergences which now form the basis of the final step in the construction of the business barometer.

c. Divergence of Time Change between Groups.—(1) Divergence between Groups.—The various groups as finally selected may move at different periods of time and in different directions at the same period of time. These diverse movements make the construction of a business barometer possible.

(2) The Group Index.—The whole value of a barometer depends upon the judicious choice of related series or groups. The selection, naturally, depends upon the purpose of the study, but for a general business index it is of prime importance to derive a group which moves in advance of general business conditions so that the lag of the general business group may be established and so furnish a forecast.

Up to this point in the discussion of a business barometer any one group may contain series of different degrees of importance and exhibit different amplitudes in their fluctuations, homogeneity alone existing in the time element of change. Therefore, the series selected to form part of any one group to be used in the barometer should be weighted according to their importance in order to obtain a group index. This may be done in a

manner similar to that discussed in the preceding chapter, the only difference being that the deviations are expressed in terms of standard deviations instead of percentages of normal; the former is a more accurate method.

(3) Forecasting.—The indices for the groups chosen for the barometer may now be plotted on the same graph. The graphic presentation should indicate the association between the groups. If group X always moves in advance of group Y it may be said that X forecasts changes in Y or Y lags behind X. In order to test this relationship the coefficient of correlation should be worked out with varying degrees of lag for the Y series, January, 1924, of X with June, 1923, of Y, etc. The association which yields the maximum coefficient of correlation may be taken as the degree of lag in Y or the period within which a change in Y will tend to follow a change in X.

In a business barometer usually more than two groups are used, but the method is similar to that given above in the derivation of the groups and their degree of association.

The importance of forecasting is apparent from the number of commercial statistical services. The *Harvard Review of Economic Statistics* uses three groups or indices: Curve A is that for speculation, B for business, and C for money. The Committee states:

> The turning points in curve A, speculation, precede, either up or down, the turning points in curve B, business, by from six to ten months throughout the entire eleven years. Similarly, the turning points in curve C, money, by from two to eight months. It is this regularity in the sequence of the movements of the three curves which affords a logical and practical basis for scientific business forecasting.[1]

164. Summary.—This chapter is but a superficial survey of the construction and use of business barometers, which make possible the forecasting of probable business conditions.

In their construction the choice of the series and their groupings are of extreme importance. By expressing cyclic fluctuations in terms of the standard deviations for the respective series a basis is found from which correlation may be measured.

By means of the coefficient of correlation the related series may be grouped into groupings, for which group indices are obtained. When these indices, by means of the coefficient of

[1] "The Harvard Index of General Business Condition, Its Interpretation" (pamphlet), p. 9.

correlation, are measured for lag, some indices may develop which may serve as a forecast of changing conditions in others.

For an admirable, and detailed discussion of the development of a business barometer the reader is referred to Review of Economic Statistics, vol. 1 by the Harvard Committee on Economic Research.

References

BABSON, ROGER W.: "Business Barometers."

COPELAND, M. T.: "Business Statistics."

DAVIES, G. R.: "An Introduction to Economic Statistics," pp. 115–116.

JORDAN, D. F.: "Business Forecasting," part II.

PERSONS, WARREN M.: "An Index of General Business Conditions," *Rev. Econ. Stat.*, vol. I, pp. 110–151, 1919.

——: "Construction of a Business Barometer Based upon Annual Data," *Amer. Econ. Rev.*, December, 1916.

——: "Correlation of Time Series," *Quar. Pub. Amer. Stat. Assoc.*, June, 1923.

VANCE, RAY: "Business and Investment Forecasting."

Problems

1. How might a business barometer be used to smooth out, to a certain extent, the crests and the troughs of the business cycle?

2. How may it be demonstrated that certain series or groups of series forecast changes in other groups?

3. Why is it necessary to use a group of series and not a single series for forecasting?

4. *a.* If a group of series is used, how may these series, though diverse in units of measurements, be made comparable?

b. What effect does (*a*) have on the amplitudes of the changes in the respective series?

c. Does (*b*) affect, in any way, the value of the series for forecasting? Explain.

5. Suppose that fifteen series had been plotted on translucent paper and that Nos. 6, 4, 3, and 2, respectively, synchronize in the time of change but differ in the degree of amplitude. To what processes would you subject the data before endeavoring to measure mathematically the degree of lag?

6. What process is used in the measurement of lag?

7. Obtain the yearly totals for business failures and immigration into the United States for the years 1900 to 1913 and work out the coefficients of correlation between the series when

a. They synchronize in time.

b. Immigration lags one year.

c. Business failures lag one year.

From your results would you say that correlation is significant and if so to what extent does one series forecast another?

8. Select another monthly series to compare with problem 11, Chap. XXIV.

From the analysis of these two series measure the degree of lag, if any, in one of the series.

9. Consult Babson's, Brookmire's, Harvard Committee on Economic Research, Moody's Investors' Service, and the Standard Statistics Company's reports for information on possible changes in business conditions and series used in forecasting.

INDEX

A

Aberthaw Co., 239
Abnormal, 35, 280, 309
Absolute (numbers), 74, 81
Accident (statistics), 17
Achenwall, G., 4
Advantages, logarithmic charts, 82
  mean, 144
  median, 152
  mode, 158
Advertising, 14
Aggregative base shifting, 263
  weighting, 265
Aids (tabulation), 47
Analysis frequency series, 133
  time series, 229
Andrew, S. L., 69
Area diagrams, 51
  preservation, 116
Arithmetic mean, assumed, 140, 168, 207
  computed, 138
  defined, 135
  in dispersion, 166
  in indices, 250, 261, 269
  in seasonal, changes, 288
  weighted, 143
Arithmetic scale in charts, 81
Array defined, 92, 93
Association, measure of, 192
Asymmetric (curves), 124
Average (deviation), 165
Averages, advantages of, 144
  attributes of, 137
  explained, 135
  for dispersion, 165
  for indices, 268, 271
  for seasonal changes, 288
  moving, 156, 287, 309
Axes (charts), 70

B

Babson's service, 13, 320
Bailey, W., 216
Baldwin, F. S., 11
Band (diagrams), 75, 76
Bar (diagrams), composite, 59
  one hundred per cent, 56
  vertical, 60
Barometer (business) construction, 321
  purpose, 230, 236, 309, 321
Base changing, 258, 263
  fixed, 257
  January, 288
  time of, 258
Bertillion, Jacques, 4
Bias, 5, 20
Bimodal, 156, 161
Block, 4
Boody, E., 308
Bowley, A. L., 4, 50, 76, 194
Bradstreets, 265
Bravais, A., 193
Brinton, W. G., 50
Brookmires Service, 13, 75, 300
Brown, W., 194
Building (index), 229
Business barometer, 230, 309, 320
  cycles, 13, 229, 320
  failures, 73
  statistics, 12

C

Captions arrangement, 45
  defined, 43
  what should be, 44
Carli, 232
Causation correlation, 193
  cyclic, 308
  seasonal, 283

Central (tendency), 133
Chaddock, R. E., 149
Chapin, Stuart, 285
Charts, field of, 64
  logarithmic, 76, 282
  necessity of, 74
  scales of, 63
  smooting, 121
  types, 40
Class interval changing, 156
  defined, 91
  width of, 94, 98
Class limits plotting, 121
  problems of, 95, 104
Classification, aids in, 33
  defined, 32
  frequency series, 91
Coefficients correlation, 198
  dispersion, 178
  regression, 222
  skewness, 190
Collection, index, 239
Collection, primary data, 23
Color (graphs), 52
Comparison by coefficients, 179, 197
  correlation, 192
  curves, 121
  cyclic changes, 317, 323
  frequency tables, 124
  ogives, 128
  mean deviation, 175
  standard deviation, 175
Component (series), defined, 41
  illustrated, 37
Composite (bar), 56
Computation of barometer, 322
  correlation, 200
  cyclic changes, 311
  deciles, 153
  deviations, mean, 165
    standard, 172
    quartile, 180
  lag, 326
  mean arithmetic, 138
    geometric, 145
    harmonic, 146
  median, 150
  mode, 158
Computation of percentiles, 153
  quartiles, 153
  regression, 218
  seasonal changes, 284
  secular trend, 296
  skewness, 187
Continuous relatives, 265, 289
  series defined, 96
    illustrated, 96
    interpolation, 126, 157
    smoothing, 116
Coonley, H., 13
Copeland, M. T., 285
Correction for correlation, 240
  deviation mean, 169
    standard, 185
  mean, 144
  secular trend, 289
Correlation and secular trend, 299
  computed, 207
  defined, 193
  limitations, 265
  perfect, 199
  positive, 193
  regression, 218
Cost (index), 143, 236
Cotsworth, Moses, 283
Crosshatching, bar, 56
  map, 62
Crum, W. L., 306
Cummings, J., 216
Cumulative (frequency), 186
Cunningham, W. J., 77
Curves, comparison of, 121
  frequency, 113
  interpretation of, 71
  logarithmic, 76
  Lorenz, 129
  ogive, 124
  staircase, 115
  superposition of, 72
  types, 122
  zone, 76
Curvilinear series, 216
  trend, 306
Cycles business, 13, 229, 297, 321
  computation, 309

D

Data, bar diagram, 58
barometer, 322
classification, 32
editing, 35
for index, 234
gathering, 29
primary, 19
secondary, 19
sources, 57
sufficiency, 18, 245
Davenport, C., 4
Davenport, E., 104
Davies, G. R., 285
Day, E. E., 37, 48, 194, 251, 257, 270
Deciles, 153
Definition of, abscissa, 64
aggregative, 263
caption, 43
classification, 32
continuous relatives, 256, 289
series, 96
correlation, 193
cycle, 308
deviation mean, 166
quartile, 180
standard, 172
discrete, 97
frequency, 41
graphs, 50
indices, 321
lag, 326
mean, arithmetic, 138
geometric, 146
harmonic, 148
median, 149
mode, 156
regression, 193
series component, 41
continuous, 96
discrete, 97
frequency, 47
spatial, 41
time, 39
skewness, 185
statistical methods, 5
statistics, 5
Definition of, stub, 43
table, 36
correlation, 195
frequency, 91
unit, 16
variations, cyclic, 326
seasonal, 284
secular, 294
Deviations in classes, 141
mean, 165
standard, 172, 307, 322
quartile, 180
Diagrams, area, 50
band, 76
bar, 56
pie, 54
Dickson, J. D., 193
Discrete (series) defined, 97
illustrated, 93
plotting, 124
smoothing, 116
Dispersion, continuous chain, 256
measures of, 165
Dot (map), 63
Douglas, P., 246, 248
Dutot, 232

E

Edgeworth, J., 4, 193
Editing, 35
Elderton, W. P., 216
Elimination of abnormal, 35
bias, 20
cyclic, 284
seasonal, 284
secular, 284
Enumeration, 24
Enumerators, duties of, 24
instructions to, 29
Error (probable), 217
Estimate (data), 25
Extremes (weight to), 144, 146, 175, 248, 258, 288

F

Federal Coal Commission, 9
Field (chart), 65

Figures, extent of use, 8
limitations of, 10
placement of, 56, 69
Fisher, Irving, 80, 81, 84, 143, 221, 250, 270, 272, 273
Fluctuations cyclic, 308
seasonal, 277
secular, 295
Forecasting, by correlation, 201, 228
by secular trend, 295
derivation of, 322
object of, 230
Formulae, arithmetic (mean), 141
coefficient of dispersion, 179
skewness, 188
correlation, 211
error, (probable), 218
"ideal," 273
mean, arithmetic, 141
geometric, 146
harmonic, 147
mean deviation, 167, 170
median, 152
mode, 158
probable (error), 218
quartiles, 154
regression, 221, 222
skewness, 188
standard (deviation), 173
variations, cyclic, 318
seasonal, 292
secular, 301, 302, 305
Frequency series analysis, 133
classes, 98
Frequency series continuous, 95
class limits, 104
defined, 41, 91
discrete, 97
graphs, 104
Frequency table, 93
total, 92

G

Galton, Sir F., 4, 193
Gehkle, C. E., 47
General tables, 37, 45
indices, 234, 245
Geometric (mean), 145, 233, 260, 263, 269
Government and statistics, 8
Graphic (methods) bar diagram, 56
correlation, 211
curves, 112
defined, 50
frequency series, 112
indeterminate (classes), 113
interpretation, 71
land diagram, 67
logarithmic, 76
Lorenz curve, 129
maps, 61
median, 152
mode, 152
ogive, 124
pie diagram, 76
quartiles, 152
regression, 223
times series, 282
cyclic, 322
seasonal, 292
secular, 297
unequal (classes), 117
zone (curve), 76
Graunt, Capt. J., 2
Growth (element), 230, 295

H

Halley, Edmund, 4
Haney, L., 21
Harmonic (mean), 147
Harvard Committee on Economic Research, 238, 245, 299, 321, 329
Hildebrand, 4
Hill, J. A., 24
History of correlation, 192
index numbers, 231
statistics, 3
Holbrook W., 306
Homogeneity, 20, 296
Hooker, R. H., 4, 217
Horizontal bar, 56
zero, 58, 110
Howland, K. E., 107

I

Ideal (formula), 253, 272
Immigration, 9, 42, 82

Index (numbers) aggregative, 263
averages to use, 268
base changing, 259
fixed, 255
period, 258
chain relative, 256
data, 234
group, 318, 327
samples for, 240
seasonal, 290
types, 236
weighting, 143, 248
Instructions, 29
Integers mid points, 98
scale, 65
Interminate (classes), 108, 113
Interpolation, 94, 102, 126, 157
Interstate Commerce Commission 9, 23
Interval class changing, 126, 157
width, 93, 98
Interval method correlation, 207
deviation mean, 168
standard, 174
mean (arithmetic), 140
Interval unequal, how to break up, 127
how to indicate, 98, 104
mean of, 142
misinterpretation of, 98

J

Jevons, W. S., 233

K

Karsten, K. G., 50, 57, 64
Kelley, T. R., 36, 193, 219, 241
King, W. I., 38, 194
Kinley, David, 14
Knies, 3
Koren, John, 5, 7

L

Lag, 326, 328
Lambertson, A., 247
Limitations of correlation, 215
line of least squares, 215, 306
mean, 144
median, 153
mode, 159
quartile deviation, 181
rectilinear charts, 81
regression, 218
secular trend, 295
Limits (class), 104
Line of equal distribution, 130
least squares, 297, 310
regression, 213
Linear (series), 216
Link (relative method), 285
Logarithmic (charts) description, 76
double, 118
differences, 79
interpretation, 84
ogive, 128
time series, 282
seasonal, 292
smoothing, 121
why used, 76
Lorenz (curve), 74, 129

M

Maps crosshatched, 61
dot, 62
pin, 63
Marriage rates, 20
Materials (collection), 29
McCann, H. K., 14
Mean arithmetic, 135, 138, 143, 166, 250, 270
deviation, 165
geometric, 145, 233, 260, 263, 269
harmonic, 147
in correlation, 202
in dispersion, 166
in indices, 258, 261, 268
monthly, 285
weighted, 143, 249, 261
Measurement (unit of) definition of, 17
indication of, 49

Median, advantages, 153
by graphic methods, 152, 227
computation, 150
defined, 135, 149
for seasonal, 288
for skewness, 189
limitations, 152
Meeker, Royal, 10
Meitzen, A., 4
Midpoint of class, 139
plotting, 112–5
rearrangement, 94, 105
variance of, 95
Miller, Douglas, 75
Mitchell, W. C., 237, 238, 240, 241, 245, 256, 259, 268
Mode advantages, 158
computation, 156
defined, 137, 156
in indices, 270
limitations, 159
Monthly data, analysis, 284, 304, 309
relatives, 286
Moody's service, 13, 320
Morbidity statistics, 10
Moving (average) 156, 297, 309

N

National Bureau of Economic Research, 9, 25, 26, 62, 98, 101, 125, 129, 131, 182, 321
National Industrial Conference Board, 59, 246
Negative correlation, 195, 200
skewness, 189
Normal curve, 123, 163, 180
distribution, 166
growth, 295, 302
production, 309, 315, 323

O

Occupational (diseases), 10
Ogive, 124
One hundred per cent graphs, 56
Ordinate defined, 64
rulings, 70
Origin (assumed) in correlation, 207
in mean, 140
in mean deviation, 169
in standard deviation, 174

P

Pareto, 131
Pearson, Karl, 4, 193
Percentage charts, 73
change, 80
deviations, 317, 323
Percentiles, 127, 153
Persons, Warren, 177, 253, 285, 321
Pètty, Sir Wm., 4
Pie (diagrams), 55, 56
Pin (map), 64
Plotting frequency curve, 112, 117
less than ogive, 124
more than ogive, 124
Polygon (frequency), 112
Population of a series, 92
of U. S. A., 37, 87
of world, 85
Positive correlation, 195, 200
skewness, 189
Prescott, Raymond, 295
Prices, relatives, 232
retail, 234, 241, 251
wholesale, 234, 242, 251, 263
Primary data, 23
tables, 37
Probable (error), 217
Production, index of, 251, 257
normal, 309
pig iron, 82
Purpose (importance), 16, 234

Q

Quadrants, negative, 209
positive, 209
Quartiles by ogive, 127
computation of, 154
deviation, 180
skewness, 189
Quetelet, 2

R

Range, 98
measure of dispersion, 165
reducing, 121
standard deviation, 176
Ratio charts, 76
of variation, 219
Ratio of change in data, 219, 299
in logarithmic charts, 78
in rectilinear charts, 78
Rectilinear (charts), 78
compared with logarithmic, 76
interpretation, 79
Regression defined, 193
lines, 218
Relationship of mean to standard deviation, 176
standard to quartile, 182
standard to range, 176
Relatives, chain, 256
continuous chain, 256, 289
link, 285
monthly, 286
representative, 286
Representative data, 23, 217
point, 128
relative, 286
Review of Economic Statistics, 13, 238, 245, 299, 308, 321, 328
Rulings, logarithmic charts, 76
rectilinear charts, 71
tables, 47

S

Sampling, 23, 217, 240
Scales, adjustment of, 74
charts, 71, 77, 112
Scatteration, coefficient of, 178
direction of, 186
measure of, 163
Schedule, 26
Seasonal causes, 283
changes, 229, 277
computation, 284
Secondary data, 19
tables, 37
Secular (trend), 277, 286, 295
Series, continuous, 95, 151
curvilinear, 216
defined, 36
discrete, 97, 114
frequency (analysis), 133
linear, 216
time (analysis), 279
types, 39, 91
Skewness defined, 185
explained, 135
measured, 187
Smoothing, continuous series, 114
ogive, 128
Snider, R. J., 237, 240, 252
Sources (data), 57
Spacing, frequency series, 101
logarithmic charts, 76
Spatial (series) defined, 41
illustrated, 41
Special indices, 234, 245
tables, 37, 45
Staircase (curve), 115
Standard (deviation) and business barometers, 322
and correlation, 198
computed, 172
defined, 170
Standards for dispersion, 166
graphs, 52
schedule, 26
secondary data, 19
tables, 48
tabulation, 48
Statistical (methods) defined, 5
Statistics defined, 5
function of, 7
history of, 3
morbidity, 10
origin, 3
population, 85, 87
Review of Economic, 13, 238, 245 257
service to business, 12
service to government, 8
Steuart, Wm., 24, 28
Stub, arrangement of, 45
choice of, 44
defined, 43
place on chart, 64

Sussmilch, J. B., 4
Synder, Carl, 252, 317

T

Tables, captions, 43
  correlation, 195
  measurements, 47
  rulings of, 47
  stub, 43
  types, 37
Tabulation, 36, 196
  machines, 34
Taxation, 10
Thorndike, E., 4
Time (series), 39
  barometers, 320
  cycles, 298
  fluctuations, 277
  seasonal change, 283
  secular trend, 277
Title for graph, 65
  for table, 42
Totals, place of, 47
Trade, indices, 239, 252
Trend, secular, 230, 277, 295
  as forecaster, 295
Types of analysis, 133
  averages, 135
  curves, 122
  data for indices, 236, 238
  fluctuations, 277
  indices, 236, 238
  samples, 240
  tables, 36

U

Unemployment, 9
Unequal (intervals) how to break up, 127
  how to illustrate, 117
  how to indicate, 98, 104
  mean of, 142
Unequal (intervals) misinterpretation of, 98
Unit defined, 17, 234, 280
  indicated, 60
U. S. Bureau of Census, 17, 24, 30, 41, 55, 57, 60, 122, 123, 129, 132
U. S. Bureau of Labor, 235, 237, 242, 244, 246, 262, 265, 272

V

Variation, cyclical, 229, 278, 308
  ratio of, 219, 299
  seasonal, 229, 277, 286
  secular, 229, 277
Verification, 35
Vertical bar diagram, 60
  scale (logarithmic), 78
  zero, 66, 112

W

Wagner, Adoph, 4
Weighting, aggregative, 265
  index numbers, 232, 248, 318, 327
  of extremes, 144, 146, 175, 288
  vitiated, 258
Wenzel, J., 66, 80, 83
Width of interval, 102

Y

Young, Arthur, 232
Yule, G. U., 4, 99, 137, 176, 193, 216, 218

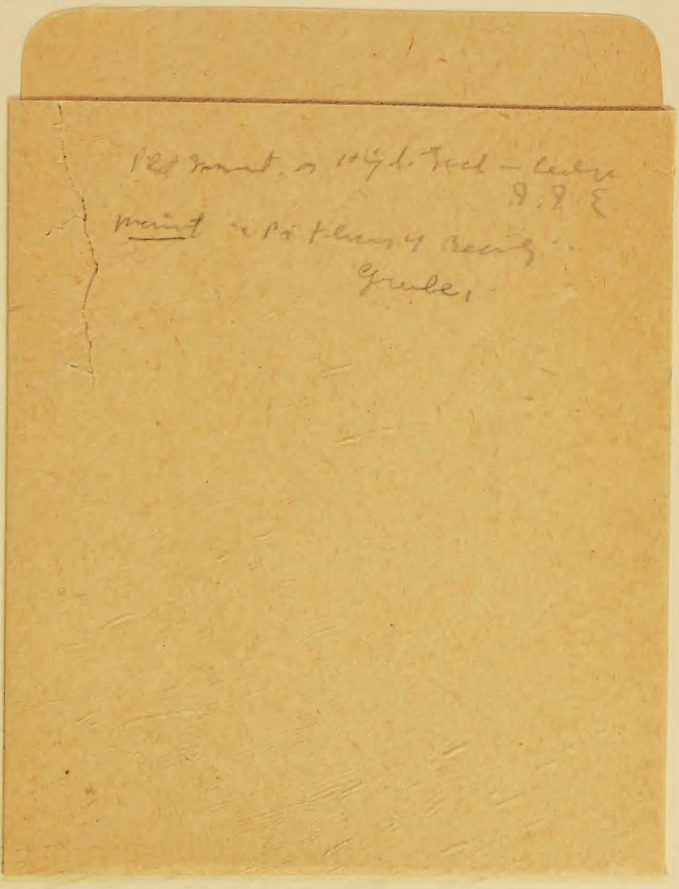
Z

Zero frequency, 100
  horizontal, 58, 112
  line in charts, 70
Zone curve, 75, 76

511
Su83

RAYMOND H. FOGLER LIBRARY
DATE DUE

NOV 14 1968			


Ref Mount. in 1/4 Std. Rod – Cadmium
A. A. E
Mount in Pitch in 4 Rings
Grade 1

UNIV OF MAINE
3 7550 01013616 7